International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 08 | Aug 2020 www.irjet.net p-ISSN: 2395-0072

TEXT DETECTION AND RECOGNITION METHODS IN DIGITAL IMAGES - A REVIEW

S.Keerthana[1], Dr.A.Suphalakshmi[2], and M.Revathi[3]

[1-3]*Department of Computer Science and Engineering, Paavai Engineering College, Namakkal, India,*

Abstract - *The emergence of machine learning and deep learning models in the field of computer vision and pattern recognition has improved the capability of Optical Character Recognition (OCR) system to recognize texts that are in arbitrary shapes, multi-orientation, and with complex backgrounds in a natural scene image or video. This paper describes the steps involved in OCR recognition, evaluation protocols, and summarizes the recent researches done in the field of text detection and text recognition.*

***Key Words*: OCR, Text Detection, Text Recognition.**

1. INTRODUCTION

Optical Character Recognition is a challenging field in computer vision and pattern Recognition which is used for digitalizing the text that is handwritten or printed within any background [1]. Without OCR the task of digitalizing the text would require recognition of text manually and typing them for long period of time.

Though Recognition of text in a controlled environment such as fixed layout, even illumination, simple background and formats has achieved greater accuracy, the recognition of text with complex layouts and backgrounds, uneven illumination in natural scene of an image or video is still a problem [2].

The applications of OCR are automatic number plate recognition, passport verification, processing bank cheques [1], digitalizing historical manuscripts, books, handwritten or typed documents, etc. They are also used in text-to-speech recognition system. OCRs are typically customized to the field of application.

Factors that influences the performance of OCR are

1) Type of the text to be recognized (typed, printed or handwritten text),
2) Text Background (simple, complex or natural scene),
3) Format of text (Image or video),
4) Language of the text (unilingual, Bilingual or Multilingual),
5) Mode of recognition (offline, online or real-time recognition).

The paper is organized as follows: Section 2 describes the steps involved in OCR, Section 3 defines the evaluation protocols used to evaluate a method, Section 4 deals with recent advancement in text detection, text recognition and end-to-end recognition methods and Section 5 concludes the paper.

2. STEPS INVOLVED IN OCR

The steps involved in OCR are

1) Data acquisition
2) Pre processing
3) Text Detection and Extraction
4) Text Enhancement
5) Text Segmentation
6) Text Recognition
7) Post processing

These steps are not delineated. They can be integrated and accomplished by using a single technique.

2.1 Data Acquisition

The data is an image of handwritten or printed text with simple, complex layouts or backgrounds in a natural scene or document. The image of the text can be obtained by

1) Digital camera:
The portability, usability and size of the digital camera make it flexible in acquiring real world text images. However the image captured are low in resolution and has uneven illumination causing blur [2].
2) Flatbed or handheld scanner:
Scanners are high in resolution, even and adequate in illumination with minimal blur and have fast batch speed. However the usability, size and portability make it difficult to capture text in natural scene images [2].
3) Datasets:
There are wide varieties of database for text images are available for researches. They are used in setting benchmarks in terms of accuracy, processing speed and storage. Some of the datasets are ICDAR datasets for text recognition in video [3], multi-lingual scene text [4], COCO-Text [5], etc. IIIT5K dataset [6] contains cropped word images, Synth90k dataset [7] contains synthetic text, CTW-1500 [8] contains curved text and Total Text [9] contains text in arbitrary shape, SVT (Street View Text) [10].

2.2 Preprocessing

Preprocessing step involves cleansing the image to improve the accuracy rates of text detection, extraction and recognition processes.

1) De-Noising:

The image of the text can be affected by sensor noise in camera, or the pixels of raw sensors that are interpolated to produce real colors, or the noises introduced by the optical scanning device such as disconnected line segments, filled loops, etc. There are several denoising methods such as Morphology based, Hough Transform based, Projection profile based, Binarization and Thresholding based, fuzzy logic based, histrogram based methods which can be used to filter the noise depending on the noise in the document image [11].

2) Document skew Detection/ Correction:

The skew angle of the document is obtained and the image is rotated accordingly. Methods used to detect skew angle of the document are profile analysis, connected components analysis, Hough transformation, etc. [12].

2.3 Text Detection and Extraction

The text detection and extraction method uses different techniques for extracting text from document and natural scene image.

Document image analysis is done to retrieve text from the document image of complex layouts, format and fonts. The major step involved in document image analysis is classification of text regions and non-text regions using text line extraction and estimation of paragraph structure, tables, etc. Current researches use deep learning models for layout analysis, document retrieval and writer identification [12][13].

Text Detection and Extraction in the natural scene image requires identifying text regions and its orientation using bounding boxes and segmenting them into appropriate groups to retain the semantic meaning of the text.

In early approaches, Connected Component Analysis (CCA) and Sliding window classification are two commonly used text extraction methods for extracting characters [14][15]. Sliding window based approaches detect characters across the image by multi-scale sliding window using a trained classifier. CC based methods segment pixels with consistent region properties such as color, edge, texture, stroke width, etc. into characters [14][15]. For text-line based method, text lines are detected initially and then each line are partitioned into multiple words [16]. The intuition behind the method is that a text region usually exhibits high self-similarity to itself and strong contrast to its local background [14].

The recent text detection methods are mostly word-based methods which fall into one of these categories: Bounding box regression based method; Segmentation based method or combined method [17].

2.4 Text Enhancement

Text Enhancement process are performed after extracting the text depending to improve the accuracy of the text recognition. The techniques involved in image Enhancement are Binarization, Slant detection and correction, text skew detection and correction. Conversion of RGB to Grayscale Image is done to reduce complexity of the problem and storage space.

- Binarization or Thresholding is a technique of converting gray scale to binary image to increase accuracy of text recognition. There are two types of thresholding techniques,
    1. Global thresholding methods that use a single threshold value to binarize the entire image. Some important global thresholding methods or fixed thresholding method are, Otsu Method, Kittler Method.
    2. Local thresholding use more than one threshold value to binarize the image. Niblack, Sauvola, Adaptive and Bersen methods are some local thresholding methods. Sauvola produces better result compared to other methods [18][19].
- Image cleansing, skew correction, line detection, slant and slope removal and character size normalization are widely used text enhancement.
- Recent methods uses Spatial Transformer Network (STN) such as Thin-plate spline transformation [20], MORN (Multi-oriented rectification network) [21], Line-fitting transformation [22] to handle text in irregular shapes and rectify them.

2.5 Text Segmentation

Segmentation is the process of segmenting the preprocessed text such as characters or words based on the technique used for text classification or recognition.
Line Segmentation:

Projection profiles are used to separate text blocks into text lines. There are two types of projection profiles: Horizontal projection profile and vertical projection profile. These techniques give better result only if the skewness in the text is corrected before analysis. There are other methods such as skeleton analysis [15].

Word or Character Segmentation:

Words or characters in the line can be segmented using vertical projection profiles analysis. However, it is difficult to estimate the optimal projection threshold for segmentation as the characters may be touching or overlapping. These problems can be solved using various adaptive and optimization methods [15].

2.6 Text Recognition

Early methods of text recognition are usually performed in a bottom-up approach, where characters are recognized initially and then they are integrated into words by using beam search, dynamic programming, etc. [14]. With the advent of deep learning models, recent approaches for word recognition tasks are considered as a multi-class classification problem [23][24], where words are categorized over a large dictionary (about 90K words, i.e., class labels) using a deep CNN. They are also solved as a sequence labelling problem where RNN or LSTM generate arbitrary length sequential labels without segmenting the characters [25], and some adopts Connectionist Temporal Classification (CTC) [26][27] to decode the sequence.

Some methods recognize text using an attention based sequence-to-sequence learning structure [16][20][21][22][28][29], where Bi-LSTMs were adopted to capture the language model from the data used for training and attention mechanism to capture the information flow within the input sequence to predict the output sequence.

To handle drifted attention problem, a FN (Focusing Network) is used to correct the drifted attention [30]. To handle misalignment caused by lost or excess character, [28] proposed Edit Probability [EP] technique.

2.7 Post Processing

In post processing step, the accuracy of the recognized text or text boundaries can be enhanced by using various loss functions or other methods.

3. EVALUATION PROTOCOLS

3.1 Text Detection

ICDAR protocols are commonly used for the evaluating the text detection methods and Word Recognition Accuracy (WRA) is used for evaluating the text recognition methods. The various evaluation protocols are overlap ratio detection protocol [31], ICDAR'03 detection protocol [32], ICDAR'11 (DelEval) detection protocol [33] and ICDAR'13 video text protocol [34] and ICDAR'17 protocols [3][4][5][35]. These various protocols are adopted for various "Robust Reading" competitions of ICDAR.

The basic measures used for text detection are precision, recall and F-measure. Precision is the ratio of the number of correctly detected text regions to the total number of detected text regions which is also known as positive predictive value (PPV). It is represented as in Eq. (1).

$$\text{Precision} = \frac{\text{No. of text regions detected correctly}}{\text{Total no. of text regions detected}} \quad (1)$$

Recall is the ratio of the number of correctly detected text regions to the total number of text regions in the dataset which is also known as sensitivity. It is represented as in Eq. (2).

$$\text{Recall} = \frac{\text{No. of text regions detected correctly}}{\text{Total no. of text regions in the dataset}} \quad (2)$$

F-score or F_1 score or f- measure is the harmonic mean of precision and recall which ranges from 0 to 1, where the best score corresponds to 1 and worst to 0. It is represented as in Eq. (3).

$$\text{f measure} = \frac{1.0}{\frac{\alpha}{\text{precision}} + \frac{1.0-\alpha}{\text{recall}}} \quad (3)$$

Where the parameter α is usually set as 0.5 to give equal importance to precision and recall.

3.2 Text Recognition

Given the cropped image of the text, the word recognition accuracy (WRA) is defined as, $\text{WRA} = \frac{|C|}{|T|}$ where C and T are number of words recognized correctly and number of ground truth labels respectively [15].

3.3 End - To - End Recognition

The text bounding box must match the ground truth and calculated using ICDAR'03 detection protocol and recognized word must match exactly as defined. Thus, the standard measures such as recognition precision, recognition recall and f-measure are used for evaluating end-to-end recognition system [15].

4. RECENT ADVANCES

The advances in text detection and recognition field over past two decades are summarized in [1][15]. This section deals with recent advances in text detection, text recognition and end-to-end recognition. Most of researches are done in detecting text that are in arbitrary shape with appropriate bounding boxes in natural scene image which are captured incidentally or intentionally.

4.1 Text Detection

Zhang et al. [36] proposed a method for detecting multi-oriented text. Two FCN were proposed. Text-block FCN that adopted VGG-16 network was used for extracting salient map from text regions in a holistic manner. The extracted salient map, extracted character component using MSER (Maximally Stable Extremal Region) and estimated orientation by component projection were used for estimating text line hypotheses. Once the text line candidates generated, it is given as input to character-centroid FCN to remove false hypotheses by predicting the centroids of the character.

Yao et al. [37] proposed a holistic approach that views text localization as semantic segmentation task. They used a Fully Convolutional Network (FCN) for estimating the information about the text regions, characters and its relationships. This method ran over the full images to produce global, pixel wise predictions maps. It adopted Holistically-Nested Edge Detection (HED) framework that performs multi-scale, multi-level feature learning to make prediction in holistic manner. From the prediction maps, texts in the natural scene were detected.

Liu and Jin [38] proposed DMPNet (Deep Matching Prior Network) that used VGG-16 as its backbone for text localization in incidental images. Quadrilateral sliding window was run over the image to remember the text regions roughly that has highly overlapping areas and shared Monte- method was used to compute areas of polygon. They proposed a sequential protocol for coordinates of the polygon to spot text with tight quadrilateral and a smooth Ln loss for text location regression.

Shi et al. [39] introduced Segment Linking (SegLink) approach for detecting oriented text. They used VGG-16 as its backbone. The texts are decomposed into segments and links. Segments are chunk of a word or text line covered by oriented box and adjacent segments are connected by links belonging to the same word or text line. Combined segments connected by links were given as detection results. The network was pre-trained with Synth-text and fine-tuned with other datasets.

He et al. [40] proposed a method based on direct regression for detecting multi-oriented texts. It is used to localize incidental scene texts in quadrilateral boundaries where identification of their characters are hard, varies in scales and perceptively distorted. The network consists of convolutional feature extraction, multi-level feature fusion and multi-task learning. Then recalled NMS, an extension of traditional NMS, was performed on the output of the network to get the detection result.

Zhou et al. [41] proposed Efficient and Accurate Scene Text Detector (EAST) which used simple and powerful pipeline for text detection. They used FCN (Fully Convolutional Network) in which PVANet (Deep but Lightweight Neural Network) was used as feature extractor and feature merging branch was used to generate numerous channels of pixel-level score map and geometry. Among the channels one was score map, where score indicates the certainty of the estimated geometry shape at a location and others were geometries enclosing the word from pixel-wise view. Two geometric shapes Rotated Box (RBOX) and Quadrangle (QUAD) were tested on text areas with separate loss functions. Then for each predicted region, Thresholding was applied and Non-Maximal Suppression (NMS) was performed to get the text detection as multi-oriented text boxes.

Jiang et al. [42] proposed Rotational Region CNN (R^2CNN), developed on Faster RCNN framework to detect arbitrarily oriented text. The RPN (Region Proposal Network) generates bounding boxes that are axis-aligned to bind the arbitrarily oriented texts and then three different ROIPoolings were performed for each generated boxes. The obtained pooled features were combined to predict text presence scores, axis-aligned and inclined box. Inclined NMS was performed on the inclined area boxes to get the detection result.

Hu et al. [43] proposed a character detector that detects any character which may be a letter in a language or a symbol in a mathematical expression. The character detector makes use of word annotations in large number of datasets to train the model by weak supervision. Once the characters were detected they undergo text structure analysis and the result was produced depending on the application.

He et al. [44] introduced a single shot detector which directly estimates bounding boxes for words in the natural scene text. They used cascaded FCN detectors in a single model which comprises of pixel-wise text supervision module, Text Attention Module (TAM) and a Hierarchical Inception Module (HIM). TAM, allows the model to use text attention map which can identify text regions roughly. HIM accumulates multi-layer inception modules, and enhances the convolutional features for the text detection task.

Dai et al. [45] proposed a framework, Fused Text Segmentation Networks (FTSN) for detecting text in multi-orientation. It comprises of feature extraction, feature fusion with region proposing and text instance prediction parts. Text detection is performed as both object detection and semantic segmentation task for detecting and segmenting the text element in parallel. It also performs a novel MNMS (Mask Non Maximum suppression) on the output of the network to produce detection result.

Zhang et al. [46] proposed a Feature Enhancement Network [FEN] for region proposing and improving text detection. The framework consists of Feature Enhancement RPN (FE-RPN) to boost text features for region proposals and Hyper Feature Generation for filtering text detection. A positive mining strategy was used on the region proposals to solve the sample Imbalance issue and adaptively weighted position-sensitive RoI pooling was used for enhancing the accuracy of the text detection.

Deng et al. [47] proposed 'PixelLink' for detecting scene text using instance segmentation. A trained CNN model was used to perform pixel-level predictions such as

Table I shows the experimental result of text detection methods on ICDAR datasets. Values are taken from the corresponding papers.

Text Detector/ Datasets	ICDAR 2011			ICDAR 2013			ICDAR 2015			ICDAR 2017 MLT		
	P	R	f (%)	P	R	f (%)	P	R	f (%)	P	R	f (%)
Zhang et al. [36]				0.88	0.78	83	0.71	0.43	54			
Yao et al. [37]				0.88	0.8	84	0.72	0.58	64.77			
TextBoxes [63]	0.89^	0.82^	86^	0.89^	0.83^	86^						
DMPNet [38]							0.73	0.68	70.64			
SegLink [39]				0.87	0.83	85.3	0.73	0.76	75			
He et al. [40]				0.92	0.81	86	0.82	0.8	81			
EAST [41]							0.83*	0.78*	80.72			
R²CNN [42]				0.93	0.82	87.73	0.85	0.79	82.54			
Wordsup [43]				0.93^	0.87^	90.34^	0.79^	0.77^	78.2^			
SSTD [44]				0.86	0.88	87	0.73	0.8	77			
FTSN [45]							0.88	0.8	84.1			
FEN [46]	**0.89^***	**0.89^***	**89.7^***	0.94^*	0.90^*	92.3^*						
PixelLink [47]				0.88*	0.87*	88.1*	0.85	0.82	83.7			
FOTS [48]						92.8^*	0.91*	0.87*	89.84*	0.81*	0.62*	70.75*
TextBoxes++ [26]				0.92^*	0.86^*	89^*	0.878*	0.785*	82.9*			
EAA [16]				0.91^	0.89^	90^	0.86^	0.87^	87^			
IncepText [48]							**0.938**	**0.873**	**90.5**			
TextSnake [49]									82.6			
Lyu et al. [51]				0.92*	0.84*	88.0*	0.89*	0.79*	84.3*	0.74*	0.70*	72.4*
SPCNet [50]				0.93	0.9	92.1	0.88	0.85	87.2	0.80*	0.68*	74.1*
MaskTextspotter [23]				0.95^	0.88^	91.7^	0.91	0.81	86			
AF-RPN [52]				0.94	0.9	92	0.89	0.83	86	0.75	0.66	70
PSENet [53]							0.88	0.85	87.08	0.77	0.68	72.45
LOMO [54]							0.87*	0.87*	87.7*	0.80*	0.67*	73.1*
PMTD [55]						93.59^	0.91	0.87	89.33	**0.84***	**0.76***	**80.13***
CRAFT [56]				**0.97^**	**0.93^**	**95.2^**	0.89	0.84	86.9	0.8	0.68	73.9
Wang et al. [17]				0.93	0.89	91.7	0.89	0.86	87.6			
Tian et al. [58]							0.88	0.85	86.6			

P- Precision, R- Recall, f- F measure. ^ indicates F-measure calculated using DelEval protocol (Wolf and Jolion, 2006), * denotes results based on multi-scale testing.

Text/non text and link prediction. Then, on applying threshold on the obtained prediction, the positive pixels were joined together by positive links, achieving instance segmentation. From the result of instance segmentation, bounding boxes of text are extracted. Post-processing was done to enhance the result.

Yang et al. [48] introduced a scene text detector named IncepText, which is based on FCIS (Fully Convolutional Instance-aware Semantic Segmentation). They used ResNet-50 as basic feature extraction module, novel Inception-Text module with deformed convolution layer to deal with text in arbitrary shapes and deformable PSROI (Position-Sensitive Region of Interest) pooling to deal with multi-oriented text detection.

Long et al. [49] introduced a flexible text representation known as TextSnake, to represent irregular text. In this method, text are represented as a number of ordered, intersecting disks, each of them located at the centre axis of text region with different radius and orientation accordingly. They used VGG-16 as its backbone in which the fully connected layers were removed and the feature maps were fed to feature merging networks at each stage of convolutions. The network produces TR (Text Region), TCL (Text Center Line) and geometry maps from which inference for text representation were obtained. It was pre-trained with Synth-text and fine-tuned with other datasets.

Xie et al. [50] proposed a text detection method developed on Mask-RCNN. They introduced a Text Context module (TCM) and a post Re-Score mechanism. The text-context module has Pyramid Attention Module (PAM) and Pyramid Fusion Module (PFM). TCM, is fed with feature maps as input and produce text segmentation as output. Then Re-Score mechanism is done to suppress FPs (False Positives).

Lyu et al. [51] proposed a combined method of the Regression based method and Segmentation based method for text detection. The network uses fully convolutional network (FCN), which consists of modules for extracting features, detecting corners and segmenting text with sensitive to position. The network outputs corner points and maps of segmentation. Corner points are sliced and grouped to generate candidate boxes, which then undergoes segmentation maps scoring and NMS suppression to give the detection result.

Zhong et al. (2018) proposed Anchor-Free Region Proposal Network (AF-RPN) for Faster R-CNN model. FPN (Feature Pyramid Network) is used as backbone and has detection modules for detecting texts of small, medium and large scales. The text proposals are obtained on thresholding score and performing NMS.

Wang et al. [53] proposed PSENet (Progressive Scale Expansion Network) for detecting arbitrary shape text in natural scene. They used Feature Pyramid Network (FPN) for obtaining feature maps which were concatenated and further fused with receptive views to get multiple segmentation masks for all text instances at certain scale. Once segmentation mask were obtained, progressive scale expansion algorithm was used to detect the arbitrary shape of the text.

Zhang et al. [54] proposed LOMO (LOok More than Once) detector for detecting arbitrary shape text. They used ResNet50 with FPN to get feature maps that was given as input to Direct Regressor (DR) to obtain quadrangle text proposals. Then Iterative Refinement Module (IRM) refines the text proposals and Shape Expression Module (SEM) reconstructs the shape of irregular text by considering geometric properties of the text elements.

Liu et al. [55] introduced Pyramid Mask Text Detector (PMTD) which was inherits Mask R-CNN and uses ResNet50 as backbone. Instead of generating a binary mask, PMTD generates a soft mask for text instances by performing regression at pixel level under location-aware supervision. Then the 2D soft mask was reinterpreted to 3D shape and a proposed plane clustering algorithm was used to derive optimal text box.

Baek et al. [56] proposed CRAFT (Character Region Awareness for Text detection) to detect text with irregularities. It uses Fully Convolutional Network (FCN), VGG-16 as backbone to give region and affinity score as designed. Region score represents the probability that given pixel, centre of the character. Affinity score represents the probability of distance between adjacent characters. The framework utilizes the character-level annotations of both synthetic and real images which are obtained by the model trained in a weak supervision.

Liu et al. [57] proposed Conditional Spatial Expansion (CSE) for curved text detection. It proved to be flexible and robust against the ambiguity caused by close texts of arbitrary orientation and suppress false positives that were included in same RoI (Region of Interest). This method progressed as region expansion process, where a seed was initialized arbitrarily in a text region then merging neighbourhood regions progressively depending on the local features extracted by a fast RCNN driven by ResNet34 and the context of information on the regions merged.

Table II shows the experimental result of text detection methods on various public datasets. Values are taken from the corresponding papers.

Text Detector/ Datasets	CTW 1500			MSRA-TD500			COCO-Text			Total-Text			RCTW-17		
	P	R	f (%)	P	R	f (%)	P	R	f (%)	P	R	f (%)	P	R	f (%)
Zhang et al. [36]				0.83	0.67	74									
Yao et al. [37]				0.76	0.75	75.91	0.43	0.27	33.31						
SegLink [39]				0.86	0.7	77									
He et al. [40]				0.77	0.7	74									
EAST [41]				0.87	0.67	76.08	0.4	0.34	37.01						
Wordsup [43]							0.45	0.3	36.8						
SSTD [44]							0.31	0.46	37						
FTSN [45]				0.87	0.77	82				0.84	0.78	81.3			
PixelLink [47]				0.83	0.73	77.8									
TextBoxes++ [26]							**0.60***	**0.56***	**58.72***						
IncepText [48]				0.87	0.79	83							0.785	0.569	66
TextSnake [49]	0.67	**0.85**	75.6			78.3				0.82	0.74	78.4			
Lyu et al. [51]				0.87	0.76	81.5	0.35*	0.34*	34.9*						
SPCNet [50]										0.83	0.82	82.9			
MaskText spotter [23]										0.69	0.55	61.3			
PSENet [53]	0.82	0.79	81.17												
LOMO [54]	0.85*	0.76*	80.8*							**0.87***	**0.79***	83.3*	**0.79***	**0.60***	**68.4***
PMTD [55]	**0.86**	0.81	**83.5**	**0.88**	0.78	82.9				**0.87**	**0.79**	**83.6**			
CRAFT [56]	0.81	0.76	78.4							0.81	**0.79**	80.2			
Wang et al. [17]	0.8	0.8	80.1	0.85	**0.82**	**83.6**				0.8	0.76	78.5			
Tian et al. [58]	0.82	0.77	80.1	0.84	0.81	82.9									

P- Precision, R- Recall, f- F measure. ^ indicates F-measure calculated using DelEval protocol (Wolf and Jolion, 2006), * denotes results based on multi-scale testing.

Wang et al. [17] proposed a detection method for detecting irregular shape scene text. It consists of Text-RPN (Region Proposal Network) that generates text proposals in initial stage and passes them to refinement stage where classification of Text/Non-text regions, bounding box regression and Recurrent Neural Network (RNN) based adaptive text region representation were performed. As the result, text regions are enclosed with polygons of flexible number of points.

Tian et al. [58] proposed segmentation based method for text detection in arbitrary shape. It uses ResNet50 as

backbone and the features were extracted from its intermediate layers. A feature merging module was used to combine the features from various layers by up-sampling and doing addition pixel-wise. Feature merging module has two branches. One generates embedding map to differentiate the text instances and other branch generates two foreground masks for text segmentation. A novel shape-aware loss and cluster processing pipeline was used as post processing to handle texts with various aspect ratio and the texts that have small gap between them. The network was pre-trained with Synth-text and fine-tuned with other datasets.

4.2 Text Recognition

Shi et al. [25] proposed Convolutional Recurrent Neural Network (CRNN), a combination of DCNN and RNN for recognizing text. In CRNN, the convolutional layers extract feature sequence and feed them to recurrent network to make predictions for per-frame of the feature sequence. The predictions are translated into a label sequence by a transcription layer. The CNN and RNN used in CRNN can be trained jointly with one loss function.

Cheng et al. [30] proposed a FAN (Focusing Attention Network) model to correct the shifted attention. It uses an attention network (AN), to recognize targeted character as same as in attention encoder and decoder model, and Focusing network (FN) to evaluate AN and perform corrections. AN utilize the features extracted from ResNet-50 based feature extractor to generate alignment points and glimpse arrays which helps FN to focus the AN on the correct target characters.

Wang and Hu [29] proposed a network named Gated RCNN (GRCNN) for recognizing text. It uses Gated Recurrent Convolution Layer (GRCL) which controls context variation in RCL, and orchestrates the feed-forward and the recurrent information. It uses Bidirectional Long Short Term Memory (BiLSTM) for modelling the sequence. The entire model of GRCNN + BLSTM can be trained end-to-end for text recognition.

Liu et al. [59] proposed a Binary Convolutional Encoder-Decoder Network (B-CEDNet) which is used along with Bidirectional Recurrent Neural Network (Bi-RNN) for recognizing text in the image. The B-CEDNet, perform character detection whereas Bi-RNN, with the contextual knowledge learned perform character level sequential correction and classification. In the salience maps produced by B-CEDNet, high confidence text region were rendered with red and white colors and from combined saliency maps a character sequence with spatial information (label vectors indicating the category, position, width and height of detected character) were extracted. The extracted sequence was fed into Bi-RNN network to perform a contextual correction and classification and to give the recognition results.

Bai et al. [28] proposed edit probability (EP), a method to improve accuracy of attention-based models for text recognition. It trains the model to handle the misalignment caused by lost or excess characters between ground truths and attention's output sequences by using its probability distribution.

Liu et al. [60] proposed a framework for text recognition where the features are learned with supervision using synthetic images. It has an image renderer, an encoder, a text decoder, an image generator, and two discriminators. The encoder extract features from an image and the text decoder estimates the character sequence from the extracted features. Synthetic images were generated by image generator. For each generated image, the parameters associated with it were obtained, and a clean image was rendered free of distortions factors, which makes it easier to be recognizable thus serving as a supervision for feature learning. The discriminators were used to improve the similarity between generated image features and clean image features.

Liao et al. [61] approached scene text recognition from a two-dimensional perspective. They proposed Character Attention Fully Convolutional Network (CA-FCN) model which uses semantic segmentation network to predict the characters at pixel level and a character attention module to highlight the foreground characters and weaken the background and also to separate adjacent characters. The 2-d character maps predicted from CA-FCN was fed to word formation module for obtaining the character sequence as result.

Zhan and Lu [22] proposed ESIR (End-to-end Scene Text Recognition via Iterative Image Rectification) with a rectification network. The network implements novel line fitting transformation that uses a polynomial to measure the centre line of texts and a pair of line segments to assess the orientation and boundary of the text lines. The evaluated transformation parameters are used by a rectification pipeline to iteratively rectify distortions. The corrected image was fed into recognition network that uses an attention-based sequence-to-sequence model for text recognition.

Luo et al. [21] proposed a Multi-Object Rectified Attention Network (MORAN) that has a Multi-Object Rectification Network (MORN) and an Attention-based Sequence Recognition Network (ASRN). MORN is trained with images and its text labels for predicting the offset for each part of the image. With the predicted offsets, sampling was applied to the original image to obtain the corrected text image. The ASRN is a framework of CNN-LSTM with an attention decoder, outputs the predictions sequentially by

Table III shows the experimental result of text recognition methods on ICDAR datasets. Values are taken from the corresponding papers.

Method/ Datasets	ICDAR 03				ICDAR 13	ICDAR 15
	50	0 (None)	Full	50k	0 (None)	0 (None)
CRNN [25]	98.7	89.4	97.6	**95.5**	86.7	
ASTER [20]	98.8	94.5	98		91.8	76.1
Jaderberg et al. [24]	98.7	93.3	**98.6**		90.8	
FAN [30]	**99.2**	94.2	97.3		93.3	70.6
GRCNN [29]	98.8	91.2	97.8			
SqueezedText [59]	98.4	93.1	97.9	93.8	92.7	
FAN [30] + EP [28]	98.7	94.6	97.9		**94.4**	73.9
Liu et al. [60]	98.1	94.7	97.5		94	
Liao et al. [61]					91.5	
ESIR [22]					91.3	**76.9**
MORAN [21]	98.7	**95**	97.8		92.4	
Baek et al. [62]		94.4			92.3	71.8

"50" and "1k" denote the lexicon used, and "None" denotes recognition without a lexicon.

Table IV shows the experimental result of text recognition methods on other datasets. Values are taken from the corresponding papers.

Method/ Datasets	SVTP	CUTE	IIIT5k			SVT	
	0 (None)	0 (None)	50	1k	0 (None)	50	0 (None)
CRNN [25]			97.6	94.4	78.2	96.4	80.8
ASTER [20]	78.5	79.5	99.6	**98.8**	**93.4**	**99.2**	**93.6**
Jaderberg et al. [24]			97.1	92.7		95.4	80.7
FAN [30]			99.3	97.5	87.4	97.1	85.9
GRCNN [29]			98	95.6	80.8	96.3	81.5
SqueezedText [59]			96.9	94.3	86.6	96.1	
FAN [30] + EP [28]			99.5	97.9	88.3	96.6	87.5
Liu et al. [60]			97.3	96.1	89.4	96.8	87.1
Liao et al. [61]		79.9	**99.8**	**98.8**	91.9	98.8	86.4
ESIR [22]	**79.6**	**83.3**	99.6	**98.8**	93.3	97.4	90.2
MORAN [21]			97.9	96.2	91.2	96.6	88.3
Baek et al. [62]	79.2	74			87.9		87.5

"50" and "1k" denote the lexicon used, and "None" denotes recognition without a lexicon.

Attending the target characters. They proposed a fractional pickup method, which was used during training phase to improve attention sensitivity.

Baek et al. [62] examined the inconsistencies in training and testing datasets which leads to variation in the performance results obtained by the existing methods. They proposed a STR (Scene Text Regconition) framework with four stages, fitting existing STR models. With that framework, an extensive evaluation was performed on recent STR modules and unexplored modules combinations. Modules are also evaluated individually over accuracy, processing speed and memory usage under common training and testing datasets. They proposed to use unique datasets for training the model as a mixture of MJSynth, SynthText and also identified the combined module of TPS +ResNet+ BiLSTM +Attention for the four stage framework produces better accuracy than the other combinations.

4.3 End-To-End Recognition

Jaderberg et al. [24] proposed an end-to-end recognition model. Though the text spotting and text recognition module are independent, the obtained recognition results were used to improve the detection results for future rounds. Region proposal method proposes number of word bounding box proposals, and a binary random forest word/no-word classifier was used to reduce the detection of false-positives. Then a CNN performs bounding box regression for cleansing the filtered proposals and another CNN was used to recognize text on each of the cleansed proposals by solving it as a classification problem upon a predefined dictionary of words. The detections results close by were merged and a rank was assigned. Non-Maximal Suppression (NMS) was performed on the detections results to obtain the final result.

Liao et al. [63] proposed a text detection method named TextBoxes and that adopted CRNN (Shi et al., 2015) for text recognition to provide end-to-end recognition model. The TextBoxes was implemented by inheriting VGG-16 appended with 9 additional convolutional layers. It also has textbox layers, linked to 6 convolutional layers for predicting a 72-dimensional vector that has text presence score (2-d) and offset (4-d) for each default text boxes (12). The outputs of the textbox layers were aggregated and NMS (Non Maximum Suppression) was performed to obtain the detection result.

Liao et al. [26] proposed TextBoxes++ for detecting arbitrary shape texts. It has inherited VGG-16 with 10 additional convolutional layers, and also has 6 textbox layers linked with 6 of the in-between convolutional layers for predicting an n-dimensional vector that has text presence scores (2-d), horizontal rectangle bounding box offsets (4-d), and rotated rectangle (5-d) or quadrilateral bounding box offsets (8-d) for each default box. The outputs of the textbox layers were aggregated and NMS (Non Maximum Suppression) was performed to obtain the

Table V shows the experimental result of end-to-end recognition methods on ICDAR-13, ICDAR 15 datasets. Values are taken from the corresponding papers.

Method/ Datasets	ICDAR 2013						ICDAR 2015					
	End-to-End			Word spotting			End-to-End			Word spotting		
	S	W	G	S	W	G	S	W	G	S	W	G
FOTS [27]	91.99*	90.11*	84.77*	95.94*	93.90*	87.76*	83.55*	79.11*	65.33*	87.01*	82.39*	67.97*
Mask TextSpotter [23]	92.2	91.1	86.5	92.5	92	88.2	79.3	73	62.4	79.3	74.5	64.2
EAA [16]	91	89	86	93	92	87	82	77	63	85	80	65
TextBoxes++ [26]	93	92	85	96	95	87	73.34	65.87	51.9	76.45	69.04	54.37
TextBoxes [63]	91	89	84	94	92	87						
ASTER [20] + TextBoxes [63]							70.6	67.3	64	75.2	71.3	67.6
Li et al. [14]							91.08	89.81	84.59	94.16	92.42	88.2
Jaderberg et al. [24]						76						

S- "Strong" lexicon provides 100 words per-image including all words that appear in the image. W- "Weak" lexicon includes all words that appear in the entire test set. And G- "Generic" lexicon is a 90k word vocabulary. * denotes results based on multi-scale testing.

detection result.

Liu et al. [27] proposed Fast Oriented Text Spotting (FOTS) model for detecting and recognizing text at the same time. The network shares the convolutional features between two tasks using to reduce the computational overhead. A proposed ROIRotate is used for generating text proposals from feature maps with respect to the estimated bounding boxes. The generated text proposals are fed to text recognition network, a framework of CNN-LSTM with CTC decoder to get the final result.

He et al. [16] proposed an end-to-end text spotter. A novel recurrent stage for recognizing text is integrated alongside with the prevailing detection branch in the CNN model used for regressing bounding box. RNN branch has a text-alignment unit that precisely computes convolutional features of a text element in irregular orientation and avoids detecting irrelevant texts and complicated background, and a LSTM-based recurrent module with a novel character attention embedding mechanism uses character spatial information as explicit supervision.

Li et al. [14] proposed a unified network for detecting and recognizing text parallelly in a single pass. The extracted convolutional features were shared for both detecting and recognizing the text thus saving the processing time. The model has CNN which outputs convolutional features and then TPN (Text Proposal network) that outputs Text proposals. The Text proposals are passed to Region Feature Encoder which gives Region Features for text detection and recognition network. The detection network output Bounding box offsets and text presence scores while the recognition network outputs recognized words.

Shi et al. [20] proposed an end-to-end text recognition model, ASTER (Attentional Scene Text Recognizer with Flexible Rectification) with rectification and recognition network. Rectification network implements novel Thin-Plate Spline (TPS) transformation to rectify irregularities in the text. Recognition network, a sequence-to-sequence model with attention mechanism and bi-directional decoder was used to predict the character sequence from the corrected image. ASTER enhances the detection mechanism by using the recognition results to filter the

Table VI shows the experimental result of end-to-end recognition methods on various datasets. Values are taken from the corresponding papers.

Method/ Datasets	ICDAR 2011	ICDAR 2017	Total-Text	SVT		SVT-50	
	Word spotting	Word spotting	End-to-End	Word spotting		Word spotting	
	G	G	W	S	G	S	G
Mask TextSpotter [23]			71.8				
TextBoxes++ [26]					64	84^	
TextBoxes [63]		87			64	84^	
Li et al. [14]	87.7			84.91	66.18		
Jaderberg et al. [24]	76				53		76

S- “Strong” lexicon provides 100 words per-image including all words that appear in the image. W- “Weak” lexicon includes all words that appear in the entire test set. And G- “Generic” lexicon is a 90k word vocabulary. ^ denotes 50 words lexicon per image.

Lyu et al. [23] proposed a new end-to-end text spotter named Mask TextSpotter based on Mask R-CNN. In this method, detecting and recognizing of text were acquired through semantic segmentation. The model has feature pyramid network (FPN) as its backbone, a region proposal network (RPN) for text proposals generation, a Fast R-CNN for regressing bounding boxes and a mask branch for instance segmentation of text and segmentation of character. It uses character-based method for text recognition and a FCN to spot and recognize characters jointly.

detection boxes and also to adjust detection boxes through rectification network.

5. CONCLUSION

The spotting of text in natural scene has wide range of applications from auto driving cars, helping tourist with language translation on recognizing text in caution board, bulletin boards, aiding visually challenged with attached text-to-speech conversion, license plate reading, etc. This literature review presents the recent models in text detection and recognition models with its performance score. From the collected results it shows that system has achieved considerable accuracy on detecting and

recognizing text that are in clear format with less complex background.

Future Scope

1) Lot of research need to be done in recognizing text in occlusion, low resolution, multi-lingual, with calligraphic fonts, and with special characters which are not covered vastly in current datasets.
2) Datasets need to be created for various languages to evaluate the broadness of a method in detecting and recognizing a text.
3) A common procedure or datasets for training and testing a method need to be created for accurate measurement of the method performance and its difference with other methods.

REFERENCES

[1] Sahu, V. and Kubde, B. (2013) 'Offline Handwritten character recognition techniques using neural network: A review', International Journal of Science and Research (IJSR), Vol. 2 No. 1, pp. 87-94.

[2] Liang, J., Doermann, D. and Li, H. (2015) 'Camera-based analysis of text and documents: a survey', International Journal on Document Analysis and Recognition, pp. 1-21.

[3] Iwamura, M., Morimoto, N., Tainaka, K., Bazazian, D., Bigorda, L.G. and Karatzas, D. (2017) 'ICDAR2017 robust reading challenge on omnidirectional video', 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), pp. 1448-1453.

[4] Nayef, N., Yin, F., Bizid, I., Choi, H., Feng, Y., Karatzas, D., Luo, Z., Pal, U., Rigaud, C., Chazalon, J. and Khlif, W. (2017) 'ICDAR2017 Robust Reading Challenge on Multi-Lingual Scene Text Detection and Script Identification - RRC-MLT', 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), pp. 1454-1459.

[5] Gomez, R., Shi, B., Bigorda, L.G., Neumann, L., Veit, A., Matas, J., Belongie, S.J. and Karatzas, D (2017) 'ICDAR2017 robust reading challenge on COCO-Text', 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), pp. 1435-1443.

[6] Mishra, A., Alahari, K. and Jawahar, C.V. (2012) 'Scene Text Recognition using Higher Order Language Priors', Proceedings of the British Machine Vision Conference, pp. 127.1– 127.11.

[7] Jaderberg, M., Simonyan, K., Vedaldi, A. and Zisserman, A. (2014) 'Synthetic data and artificial neural networks for natural scene text recognition', ArXiv, abs/1406.2227.

[8] Liu, Y., Jin, L., Zhang, S. and Zhang, S. (2017) 'Detecting curve text in the wild: New dataset and new solution', ArXiv, abs/1712.02170.

[9] Ch'ng, C.K. and Chan, C.S. (2017) 'Total-Text: A Comprehensive Dataset for Scene Text Detection and Recognition', 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), pp. 935-942.

[10] Wang, K., Babenko, B. and Belongie, S.J. (2011) 'End-to-end scene text recognition', International Conference on Computer Vision, pp. 1457-1464.

[11] Farahmand, A., Sarrafzadeh, A. and Shanbehzadeh, J. (2013) 'Document Image Noises and Removal Methods', proc. of the International MultiConference of Engineers and Computer Scientists, pp. 436-440.

[12] Moysset, B., Kermorvant, C. and Wolf, C. (2018) 'Learning to detect, localize and recognize many text objects in document images from few examples', International Journal on Document Analysis and Recognition, Vol. 21 No.3, pp. 161–175.

[13] Liu, CL., Fink, G.A., Govindaraju, V. and Jin, L. (2018) 'Special issue on deep learning for document analysis and recognition', International Journal on Document Analysis and Recognition (IJDAR), Vol. 21 No. 3, pp. 159-160.

[14] Li, H., Wang, P. and Shen, C. (2017) 'Towards End-to-End Text Spotting with Convolutional Recurrent Neural Networks', 2017 IEEE International Conference on Computer Vision (ICCV), pp. 5248-5256.

[15] Ye, Q. and Doermann, D. (2015) 'Text Detection and Recognition in Imagery: A Survey', IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 37 No. 7, pp. 1480-1500.

[16] He, T., Tian, Z., Huang, W., Shen, C., Qiao, Y. and Sun, C. (2018) 'An end-to-end TextSpotter with Explicit Alignment and Attention', IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5020-5029.

[17] Wang, X., Jiang, Y., Luo, Z., Liu, C., Choi, H. and Kim, S. (2019) 'Arbitrary Shape Scene Text Detection with Adaptive Text Region Representation', IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6442-6451.

[18] Bahi, H.E., Mahani, Z. and Zatni, A. (2014) 'An offline handwritten character recognition system for image obtained by camera phone', 19th International Conference on Applied Mathematics, Istanbul, Turkey, pp. 180-189.

[19] Puneet and Garg, N. (2013) 'Binarization techniques used for grey scale images', International Journal of Computer Applications, Vol. 71 No. 1, pp. 8-11.

[20] Shi, B., Yang, M., Wang, X., Lyu, P., Yao, C. and Bai, X. (2018) 'ASTER: An Attentional Scene Text Recognizer with Flexible Rectification', IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 41 No. 9, pp. 2035-2048.

[21] Luo, C., Jin, L. and Sun, Z. (2019) 'MORAN: A Multi-Object Rectified Attention Network for scene text recognition', Pattern Recognition, Vol. 90, pp. 109-118.

[22] Zhan, F. and Lu, S. (2018) 'ESIR: End-to-end Scene Text Recognition via Iterative Image Rectification', 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2054-2063.

[23] Lyu, P., Liao, M., Yao, C., Wu, W. and Bai, X. (2018) 'Mask TextSpotter: An End-to-End Trainable Neural Network for Spotting Text with Arbitrary Shapes', Computer Vision – European Conference on Computer Vision 2018, pp 71-88.

[24] Jaderberg, M., Simonyan, K., Vedaldi, A. and Zisserman, A. (2014) 'Reading Text in the Wild with Convolutional Neural Networks', International Journal of Computer Vision (IJCV), Vol. 116, pp. 1-20.

[25] Shi, B., Bai, X. and Yao, C. (2015) 'An End-to-End Trainable Neural Network for Image-Based Sequence Recognition and Its Application to Scene Text Recognition', IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 39 No. 11, pp. 2298–2304.

[26] Liao, M., Shi, B. and Bai, X. (2018) 'Textboxes++: A single-shot oriented scene text detector', IEEE transactions on image processing, Vol. 27 No. 8, pp. 3676-3690.

[27] Liu, X., Liang, D., Yan, S., Chen, D., Qiao, Y. and Yan, J. (2018) 'FOTS: fast oriented text spotting with a unified network', IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5676-5685.

[28] Bai, F., Cheng, Z., Niu, Y., Pu, S. and Zhou, S. (2018) 'Edit Probability for Scene Text Recognition', IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1508-1516.

[29] Wang, J. and Hu, X. (2017) 'Gated Recurrent Convolution Neural Network for OCR', Proceedings of Neural Information Processing Systems (NIPS), pp. 334–343.

[30] Cheng, Z., Bai, F., Xu, Y., Zheng, G., Pu, S. and Zhou, S. (2017) 'Focusing Attention: Towards Accurate Text Recognition in Natural Images', 2017 IEEE International Conference on Computer Vision (ICCV), pp. 5086-5094.

[31] Hua, X., Liu, W. and Zhang, H. (2004) 'An automatic performance evaluation protocol for video text detection algorithms', IEEE Transactions on Circuits and Systems for Video Technology, Vol.14 No.4, pp. 498-507.

[32] Lucas, S.M., Panaretos, A., Sosa, L., Tang, A., Wong, S. and Young, R. (2003) 'ICDAR 2003 robust reading competitions', Proceedings of the 7th International Conference on Document Analysis and Recognition, pp. 682-687.

[33] Wolf, C. and Jolion, J. (2006) 'Object Count / Area Graphs for the Evaluation of Object Detection and Segmentation Algorithms', International Journal of Document Analysis and Recognition (IJDAR), Vol. 8 No. 4, pp. 280-296.

[34] Karatzas, D., Shafait, F., Uchida, S., Iwamura, M., Bigorda, L.G., Mestre, S.R., Romeu, J.M., Mota, D.F., Almazán, J. and Heras, L.D. (2013) 'ICDAR 2013 Robust Reading Competition', 12th International Conference on Document Analysis and Recognition, pp. 1484-1493.

[35] Yang, C., Yin, X., Yu, H., Karatzas, D. and Cao, Y. (2017) 'ICDAR2017 robust reading challenge on text extraction from biomedical literature figures (DeTEXT)', 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), Vol. 1, pp. 1444-1447.

[36] Zhang, Z., Zhang, C., Shen, W., Yao, C., Liu, W. and Bai, X. (2016) 'Multi-Oriented Text Detection with Fully Convolutional Networks', IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4159-4167.

[37] Yao, C., Bai, X., Sang, N., Zhou, X., Zhou, S. and Cao, Z. (2016) 'Scene Text Detection via Holistic, Multi-Channel Prediction', ArXiv, abs/1606.09002.

[38] Liu, Y. and Jin, L. (2017) 'Deep Matching Prior Network: Toward Tighter Multi-oriented Text Detection', IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3454-3461.

[39] Shi, B., Bai, X. and Belongie, S.J. (2017) 'Detecting Oriented Text in Natural Images by Linking Segments', IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3482-3490.

[40] He, W., Zhang, X., Yin, F. and Liu, C. (2017) 'Deep direct regression for multi-oriented scene text detection', IEEE International Conference on Computer Vision (ICCV), pp. 745-753.

[41] Zhou, X., Yao, C., Wen, H., Wang, Y., Zhou, S., He, W. and Liang, J. (2017) 'EAST: An Efficient and Accurate Scene Text Detector', IEEE Conference on Computer Vision and Pattern Recognition, pp. 5551-5560.

[42] Jiang, Y., Zhu, X., Wang, X., Yang, S., Li, W., Wang, H., Fu, P. and Luo, Z. (2017) 'R^{2}CNN: Rotational Region CNN for Orientation Robust Scene Text Detection', ArXiv, abs/1706.09579.

[43] Hu, H., Zhang, C., Luo, Y., Wang, Y., Han, J. and Ding, E. (2017) 'WordSup: Exploiting Word Annotations for Character based Text Detection', IEEE International Conference on Computer Vision (ICCV), pp. 4950-4959.

[44] He, P., Huang, W., He, T., Zhu, Q., Qiao, Y. and Li, X. (2017) 'Single Shot Text Detector with Regional Attention', IEEE International Conference on Computer Vision (ICCV), pp. 3066-3074.

[45] Dai, Y., Huang, Z., Gao, Y., Xu, Y., Chen, K.P., Guo, J. and Qiu, W. (2018) 'Fused Text Segmentation Networks for Multi-oriented Scene Text Detection', 24th International Conference on Pattern Recognition (ICPR), pp. 3604-3609.

[46] Zhang, S., Liu, Y., Jin, L. and Luo, C. (2017) 'Feature Enhancement Network: A Refined Scene Text Detector', ArXiv, abs/1711.04249.

[47] Deng, D., Liu, H., Li, X. and Cai, D. (2018) 'Pixellink: Detecting scene text via instance segmentation', ArXiv, abs/1801.01315.

[48] Yang, Q., Cheng, M., Zhou, W., Chen, Y., Qiu, M. and Lin, W. (2018) 'IncepText: A New Inception-Text Module with Deformable PSROI Pooling for Multi-Oriented Scene Text Detection', Proceedings of the 27th International Joint Conference on Artificial Intelligence, pp. 1071–1077.

[49] Long, S., Ruan, J., Zhang, W., He, X., Wu, W. and Yao, C. (2018) 'TextSnake: A Flexible Representation for Detecting Text of Arbitrary Shapes' in European Conference on Computer Vision, pp. 19-35.

[50] Xie, E., Zang, Y., Shao, S., Yu, G., Yao, C. and Li, G. (2018) 'Scene Text Detection with Supervised Pyramid Context Network', ArXiv, abs/1811.08605.

[51] Lyu, P., Yao, C., Wu, W., Yan, S. and Bai, X. (2018) 'Multi-oriented Scene Text Detection via Corner Localization and Region Segmentation', IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7553-7563.

[52] Zhong, Z., Sun, L. and Huo, Q. (2018) 'An Anchor-Free Region Proposal Network for Faster R-CNN based Text Detection Approaches', International Journal on Document Analysis and Recognition (IJDAR), Vol. 22 No. 3, pp. 315–327.

[53] Wang, W., Xie, E., Li, X., Hou, W., Lu, T., Yu, G. and Shao, S. (2019) 'Shape Robust Text Detection with Progressive Scale Expansion Network', IEEE Conference on Computer Vision and Pattern Recognition, pp. 9336–9345.

[54] Zhang, C., Liang, B., Huang, Z., En, M., Han, J., Ding, E. and Ding, X. (2019) 'Look More Than Once: An Accurate Detector for Text of Arbitrary Shapes', IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.10544-10553.

[55] Liu, J., Liu, X., Sheng, J., Liang, D., Li, X., & Liu, Q. (2019) 'Pyramid Mask Text Detector', arXiv preprint arXiv: 1903.11800.

[56] Baek, Y., Lee, B., Han, D., Yun, S. and Lee, H. (2019) 'Character Region Awareness for Text Detection', IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9357-9366.

[57] Liu, Z., Lin, G., Yang, S., Liu, F., Lin, W. and Goh, W.L. (2019) 'Towards Robust Curve Text Detection with Conditional Spatial Expansion', IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7261-7270.

[58] Tian, Z., Shu, M., Lyu, P., Li, R., Zhou, C., Shen, X. and Jia, J. (2019) 'Learning Shape-Aware Embedding for Scene Text Detection', IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4229-4238.

[59] Liu, Z., Li, Y., Ren, F., Goh, W.L. and Yu, H. (2018) 'SqueezedText: A Real-Time Scene Text Recognition by Binary Convolutional Encoder-Decoder Network', 32nd AAAI Conference on Artificial Intelligence, AAAI 2018, pp. 7194-7201.

[60] Liu, Y.P., Wang, Z., Jin, H. and Wassell, I.J. (2018), 'Synthetically Supervised Feature Learning for Scene Text Recognition', European Conference on Computer Vision (ECCV), pp. 449-465.

[61] Liao, M., Zhang, J., Wan, Z., Xie, F., Liang, J., Lyu, P., Yao, C. and Bai, X. (2019) 'Scene Text Recognition from Two-Dimensional Perspective', Proceedings of the 33rd AAAI Conference on Artificial Intelligence, pp. 8714–8721.

[62] Baek, J., Kim, G., Lee, J., Park, S., Han, D., Yun, S., Oh, S.J. and Lee, H. (2019) 'What is wrong with scene text recognition model comparisons? dataset and model analysis', ArXiv, abs/1904.01906.

[63] Liao, M., Shi, B., Bai, X., Wang, X. and Liu, W. (2016) 'TextBoxes: A Fast Text Detector with a Single Deep Neural Network', ArXiv, abs/1611.06779.