A Survey on Multi Oriented Text Recognition

Nisarg Gandhewar[1], S. R. Tandan[2]
[1]S. B. Jain Institute of Technology, Management & Research, Nagpur (M.S.) India.
[2]Dr. C.V. Raman University, Bilaspur (C.G.) India.

ABSTRACT
Increasing use of smart phone in our day to day life to capture images initiates a need to recognize text from natural images which is nowadays a hot research topic in the field of computer vision due to its various applications. Text in natural scenes exists in almost every phase of our daily life. From the facade of the buildings in our city to the cover of a book in our library. Undoubtedly, text is among the most brilliant and influential creations of humankind. Despite the enduring research of several decades on optical character recognition (OCR), recognizing texts from natural images is still a difficult task because of series of grand challenges which is still be encountered when detecting and recognizing text. Scene texts are often found in irregular shape (curved, or arbitrarily oriented) and its recognition not yet been well addressed in the literature. Most of the existing methods on text recognition work with regular (horizontal) texts and not generalized to handle irregular texts. This survey is aimed at summarizing and analyzing the major changes and significant progress of multi oriented text recognition in the deep learning era.

Keywords: Scene text recognition, optical character recognition, deep learning. Smart phone.

I INTRODUCTION

Scene text recognition is an essential process in computer vision tasks. Many practical applications such as traffic sign reading, product recognition, intelligent inspection, and image searching, benefit from the rich semantic information of scene text. With the development of scene text detection methods, scene character recognition has emerged at the forefront of this research topic and is regarded as an open and very challenging research problem.

Nowadays, regular text recognition methods have achieved notable success. More-over, methods based on convolution neural networks have been broadly applied. Integrating recognition models with recurrent neural networks and attention mechanisms yields better performance for these models. Nevertheless, most current recognition models remain too unstable to handle multiple disturbances from the environment. Furthermore, the various shapes and distorted patterns of irregular text cause additional challenges in recognition. As illustrated in Fig. 1, scene text with irregular shapes, such as perspective and curved text, is still very challenging to recognize.(Liu et al. 2019; Luo et al. 2019),Curved text detection is a difficult problem that has not been addressed sufficiently.

Reading text is naturally regarded as a multi classification task involving sequence-like objects. Usually, the characters in one text are of the same size. However, characters in different scene texts can vary in size.

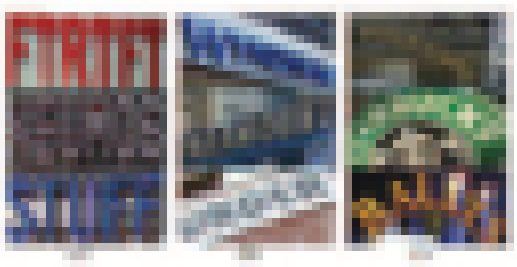

Fig No.1 Examples of regular and irregular scene text. (a)Regular text. (b) Slanted and perspective text. (c) Curved text.

II REVIEW OF LITERATURE

Scene text understanding essentially includes two tasks: text detection and word recognition. Here we present a brief introduction to related works on text detection, word recognition, and text spotting systems that combine both.

(a) Text Detection:
Text detection aims to localize text in images and generate bounding boxes for words. Existing approaches can be roughly classified into three categories: character based, text-line based and word based methods (Li et al, 2017).

(i) **Character based methods:** It firstly finds characters in images, and then group them into words. They can be further divided into sliding window based (Jaderberg M et al, 2014; Wang T et al., 2018) and Connected Components (CC) based methods. Sliding window based approaches use a trained classifier to detect characters across the image in a multi-scale sliding window fashion.CC based methods segment pixels with consistent region properties (i.e., color, stroke width, density, etc.) into characters. The detected characters are further grouped

into text regions by morphological operations, conditional random fields or other graph models.

(ii) **Text-line based methods:** It detects text lines firstly and then separates each line into multiple words. The motivation is that people usually distinguish text regions initially even if characters are not recognized. Based on the observation that text region usually exhibits high self-similarity to itself and strong contrast to its local background, (Zhang et al, 2016)propose to extract text lines by exploiting symmetry property and localize text lines via salient maps that are calculated by fully convolution networks. Post processing techniques are also proposed in to extract text lines in multiple orientations.

(iii) **Word Based Methods**: More recently, a number of approaches are proposed to **detect words** directly using DNN based techniques, such as Faster R-CNN, YOLO, SSD (Li et al, 2017). By extending Faster R-CNN, (Zhong et al. 2017) design a text detector witha multi-scale Region Proposal Network (RPN) and a multi-level ROI pooling layer. (Tian et al., 2016) Develop a vertical anchor mechanism, and propose a Connectionist Text Proposal Network (CTPN) to accurately localize text lines in images. (Gupta et al. 2016) use a Fully Convolution Regression Network (FCRN) for efficient text detection and bounding box regression, motivated by YOLO. Similar to SSD, (Liao et al. 2017) propose "Text Boxes" by combining predictions from multiple feature maps with different resolutions, and achieve the best-reported text detection performance on datasets.

(b) Text Recognition: Traditional approaches to text recognition usually perform in a bottom-up fashion, which recognize individual characters firstly and then integrate them into words by means of beam search, dynamic programming, etc. In contrast, (Jaderberg et al.2014) consider word recognition as a multi-class classification problem, and categorize each word over a large dictionary (about 90Kwords) using a deep convolutional neural network (CNN).With the success of RNNs on and writing recognition, He et al. and Shi et al. treat word recognition as a sequence labeling problem. RNNs are employed to generate sequential labels of arbitrary length without character segmentation, and Connectionist Temporal Classification (CTC) is adopted to decode the sequence. (Shi X, et al, 2016) propose to recognize text using an attention-based sequence-to-sequence learn-ing structure. In this manner, RNNs automatically learn the character-level language model presented in word strings from the training data. The soft-attention mechanism allows the model to selectively exploit local image features. These networks can be trained end-to-end with cropped word patches as inputs. Moreover, Shi et al. insert a Spatial Transformer Network (STN) to handle words with irregular shapes.

(i) **Multi-Oriented Text Recognition:** Compared with regular text recognition work, irregular text recognition is more difficult. One kind of irregular text recognition method is the bottom-up approach which searches for the position of each character and then connects them. Another is the top-down approach. This type of approach matches the shape of the text, attempts to rectify it, and reduces the degree of recognition difficulty.

In the bottom-up manner, a two-dimensional attention mechanism for irregular text was proposed by (Yang et al. 2017)Based on the sliced Wasserstein distance, the attention alignment loss is adopted in the training phase, which enables the attention model to accurately extract the character features while ignoring the redundant background information. (Cheng et al. 2018), proposed an arbitrary-orientation text recognition network, which uses more direct information of the position to instruct the network to identify characters in special locations.

In the top-down manner, STAR-Net used an affine transformation network that transforms the rotated and differently scaled text into more regular text. Meanwhile, a ResNet is used to extract features and handle more complex background noise. RARE regresses the fiducially transformation points on sloped text and even curved text, thereby mapping the corresponding points onto standard positions of the new image. Using thin plate spline to back propagate the gradients, RARE is end-to-end optimized. (Luo et al. 2019) proposed a MORAN model which uses the top-down approach. The MORAN consists of a multi-object rectification network and an attention-based sequence recognition network. The multi-object rectification network is designed for rectifying images that contain irregular text. It decreases the difficulty of recognition and enables the attention-based sequence recognition network to more easily read irregular text. But the MORAN will fail when the curve angle in text is too large.

(c) Text Spotting: Text spotting needs to handle both text detection and word recognition. (Wang et al. 2017) take the locations and scores of detected characters as input and try to find an optimal configuration of a particular word in a given lexicon, based on a pictorial structures formulation. (Neumann et al. 2013) use a CC based method for character detection. These characters are then agglomerated into text lines based on heuristic rules. Optimal sequences are finally found in each text line using dynamic programming, which are the recognized words. These recognition-based pipelines lack explicit

word detection. Some text spotting systems firstly generate text proposals with a high recall and a low precision, and then refine them using a separate recognition model it is expected that a strong recognizer can reject false positives, especially when a lexicon is given. (Jaderberg et al. 2014) use an ensemble model to generate text proposals, and then adopt the word classifier in for recognition. (Gupta et al. 2016) em-ploy FCRN for text detection and the word classifier in for recognition. (Liao et al. 2017) combine "Text Boxes" and "CRNN", which yield state-of-the-art text spotting performance on datasets.

III BENCHMARK DATASETS

Despite the success in Scene text detection & recognition, current methods are only evaluated on single datasets after being trained on them separately. Naturally, a new dataset representing several challenges would also provide extra momentum for this field. Evaluation of cross dataset generalization ability is also preferable, where the model is trained only on one dataset and then tested of another. We have collected existing datasets supporting multi oriented text summarized in Tab.1.

The CTW1500 dataset contains 1500 images, with 10,751 bounding boxes (3,530 are curved bounding boxes) and at least one curved text per image. The images were manually extracted from the Internet, image libraries, such as Google Open-Image, and data collected via phone cameras, which also contain a large amount of horizontal and multi-oriented text. The distribution contains indoor, outdoor, born digital, blurred, perspective distortion text and other text. This dataset is multilingual, containing mostly Chinese and English text. Some of the images from this dataset is shown in fig 6.

Table 1
Datasets with support for multi oriented text

Data Set	Images Training / Testing	Orientation	Language	Remark
CTW1500 (2019)	1000/500	Multioriented	English, Chinese	
ICDAR 2015	1000/500	Multioriented	English	Blur Images
ICDAR RCTW(2017)	8034/4229	Multioriented	Chinese	
Total-Text (2017)	1255/300	Curved	English, Chinese	
CTW (2017)	25K/6K	Multioriented	Chinese	
COCO-TEXT (2017)	63686/43686	Multioriented	English	Some images do not contain text
MSRA-TD500	300/200	Multioriented	English, Chinese	long text

In the same way Fig 2, Fig 3, Fig 4 and Fig 5 shows sample images from MSRA-TD500, ICDAR2015, ICDAR2013, COCO TEXT dataset. (Liu Y, et al. 2019;Ma J et al. 2018),

Fig No.2 MSRA-TD500

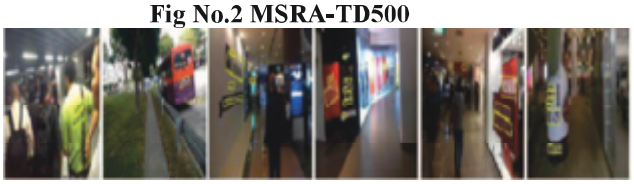

Fig No.3 ICDAR2015

Fig No.4 ICDAR2013

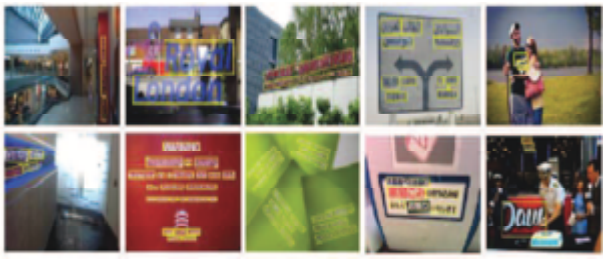

Fig No.5. Examples of detection results. From left to right in columns: ICDAR2015, ICDAR2013, MSRA-TD500, MLT, COCO-Text.

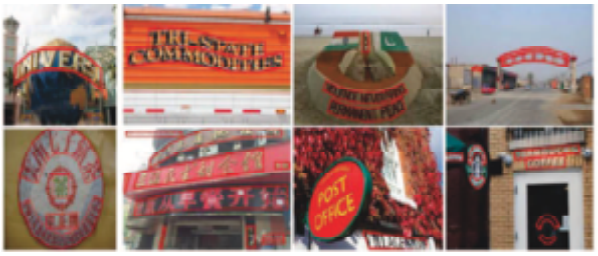

Fig No.6 Examples of annotations in the CTW1500 dataset [1]

IV SIGNIFICANT WORK FOR MULTI ORIENTED TEXT RECOGNITION

Table 2
Significant Work on Multi Oriented Text Recognition

Sr No	Author Name	Year	Proposed Work	Dataset Used	Remark
1	Yuliang Liu et al [1]	5 Feb 19	(Liu et al. 2019) proposed a new dataset called CTW1500, comprising mainly English and Chinese curved text. They also proposed a polygon-based curved text detector that can detect curved text without using an empirical combination	CTW1500, Total-text, MSRA-TD500	Proposed dataset could be enlarged as a curved-text-based recognition dataset
2	Canjie Luo et al. [2]	10 Jan 19	(Luo et al. 2019) Proposed a multi-object rectified attention network (MORAN) for scene text recognition. The proposed framework involves two stages: rectification and recognition. rectification transform an image containing irregular text into a more readable One. For recognition attention-based sequence recognition network was designed to recognize the rectified image and outputs the characters in sequence	IIIT5K, SVT, ICDAR2003, ICDAR2013, ICDAR2015, SVT-Perspective and CUTE80	The MORAN will fail when the curve angle is too large
3	Zhanzhan Cheng et al. [3]	22 Mar 18	(Cheng et al. 2018) develop the arbitrary orientation network (AON) to directly capture the deep features of irregular texts, which are combined into an attention-based decoder to generate character sequence. develop an arbitrary orientation network consisting of the horizontal network (HN), the vertical network (VN) and the character placement clue network (CN) for extracting horizontal, vertical and placement features respectively.	CUTE80, SVT-Perspective, IIIT5k, SVT and ICDAR	Computational cost and the number of parameters are major concerns in resource-constrained scenarios such as embedded computer Systems.
4	Xiao Yang et al [4]	19 Aug 17	(Yang et al. 2017) Developed a robust end-to-end neural-based model which includes two learning components: (1) an auxiliary dense character detection task using a FCN that helps to learn text specific visual patterns, (2) an alignment loss that provides guidance to the training of an attention model. Also generated a large-scale synthetic dataset containing perspectively distorted and curved text.	SVT-Perspective ,CUTE80, ICDAR03, III5K	Future directions would be to combine the proposed text recognition model with a text detection method for a full end-to-end system.

5	Jianqi Ma et al[5]	15 Mar 18	(Ma et al. 2018) present the Rotation Region Proposal Networks (RRPN), which are designed to generate inclined proposals with text orientation angle information. The angle information is then adapted for bounding box regression to make the proposals more accurately fit into the text region in terms of the orientation.	MSRA-TD500, ICDAR 2015, ICDAR 2013,	consider only three datasets
6	Pengyuan Lyu et al [6]	27 Feb 18	(Lyu et al. 2018)propose a model to detect scene text by localizing corner points of text bounding boxes and segmenting text regions in relative positions. In inference stage, candidate boxes are generated by sampling and grouping corner points, which are further scored by segmentation maps and suppressed by NMS.	ICDAR2013, ICDAR2015, MSRA-TD500, MLT and COCO-Text	When two text instances are extremely close, it may predict the twoinstances as one.
7	Baoguang Shi et al [7]	12 Mar 16	(Shi et al 2016), propose RARE (Robust text recognizer with Automatic rectification) RARE is a specially designed deep neural network, which consists of a Spatial Transformer Network (STN) and a Sequence Recognition Network (SRN)	IIIT 5K-Words,Street View Text, ICDAR 2003, ICDAR 2013	It did not address the end-to-end scene text reading problem.

V CONCLUSION

The past several years have witness the significant development of methods for detecting & recognizing a text. Despite the success so far, methods for text detection and recognition are still confronted with several challenges. Existing methods on text recognition mainly work with regular (horizontal) texts and not generalized to handle irregular texts. While human have barely no difficulties localizing and recognizing text, current methods are not designed and trained effortlessly.

The recent work on multi oriented text recognition address the problem of handling irregular text but still they have not yet reached human-level performance & still there is scope for improvement. Few datasets are available which supports irregular text but still we require more to train and evaluate current and future methods efficiently.

REFERENCES

[1] Liu Y, Jin L, (2019), Curved scene text detection via transverse and longitudinal sequence connection, *Elsevier Journal of Pattern Recognition* (Vol 90), PP 337-345.

[2] Luo C, Lianwen J, (2019), MORAN: A Multi-Object Rectified Attention Network for Scene Text Recognition, *Journal of Pattern Recognition, Science Direct, (*Vol 90), Pages 109-118.

[3] Cheng Z, Bai F, (2018), AON: Towards arbitrarily-oriented text recognition, *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5571–5579.

[4] Yang X, He D, (2017), Learning to Read Irregular Text with Attention Mechanisms, Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence.

[5] Ma J, Shao W, (2018), Arbitrary-Oriented Scene Text Detection via Rotation Proposals, Draft submitted to cornell university.

[6] Lyu P ,Yao C, (2018), Multi-Oriented Scene Text Detection via Corner Localization and Region Segmentation, Draft submitted to Cornell University.

[7] Shi B, Wang X, (2016), Robust Scene Text Recognition with Automatic Rectification, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*

[8] Li Hui, Wang P, (2017), Towards End-to-end Text Spotting with Convolutional Recurrent Neural Networks *IEEE International Conference on Computer Vision (ICCV).*

[9] Q. Ye and. Doermann D, (2015), Text detection and recognition in imagery: A survey, *IEEE Trans. Pattern Anal. Mach. Intell.,37(7):1480–1500.*

[10] Zhu Y, Yao C, (2016), Scene text detection and recognition: recent advances and future trends, *Frontiers of Computer Science, 10(1):19–36.*

[11] Jaderberg M, Vedaldi A,(2014), Deep features for text spotting, *In Proc. European. Conference of Computer Vision.*

[12] Wang T, (2012), End-to-end text recognition with convolutional neural networks, *In Proc. IEEE Int. Conf. Pattern. Recognition.*

[13] Zhu S, Zanibbi R,(2016), A text detection system for natural scenes with convolutional feature learning and cascaded classification, *In Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*

[14] Huang W, Lin Z, (2013), Text localizationin natural images using stroke feature transform and text co-variance descriptors, *In Proceeding IEEE International. Conference. Computer. Vision.*

[15] Zhang, Z., Zhang, C., Shen, W., Yao,C., and Bai,X. (2016), Multi-oriented text detection with fully convolutional networks", In Proc. IEEE Conf. Comp. Vis. Patt. Recogn.

[16] Tian Z.,Huang W.,He T.,He P., and Qiao Y., (2016), Detecting text in natural image with connectionist text proposal network", In Proc. Eur. Conf. Comp. Vis..

[17] Neumann L, Matas J, (2013), Scene text localization and recognition with oriented stroke detection, In Proc. IEEE Int. Conf. Comp. Vis., 2013.

[18] Gupta A, Zisserman A, (2016), Synthetic data for text localisation in natural images", In Proc. IEEE Conf. Comp. Vis. Patt. Recogn.

[19] Liao M, Shi B, Bai B, Wang X, and Liu W, (2017) Textboxes: A fast text detector with a single deep neural network, In Proc.National Conf. Artificial Intell.

[20] Tian, Z, Huang W, He T, He P, and Qiao Y,(2016), Detecting text in natural image with connectionist text proposal network, In Proc. Eur. Conf. Comp. Vis.