January - February 2020
ISSN: 0193-4120 Page No. 16917 - 16930

Comprehensive Study on Recent Advances and Future Trends in Detection & Recognition of Scene Text

Nisarg Gandhewar[1], Dr S R Tandon[2], Dr Rohit Miri [3]
[1]PhD Scholar,[2,3]Associate professor

Article Info
Volume 82
Page Number: 16917 - 16930
Publication Issue:
January - February 2020

Article History
Article Received*: 18 May 2019*
Revised*: 14 July 2019*
Accepted*: 22 December 2019*
Publication*: 29 February 2020*

Abstract

Increasing use of smart phone in our daily life to capture images initiates a need to recognize text from natural images which is nowadays a promising research topic in the field of computer vision and document analysis due to its various applications. Text in natural scenes exists in almost every phase of our daily life. From the front of the buildings in our locality to the cover of a book in library. Certainly, text is one of the most radiant and powerful creations of human kind. Despite of variety of challenges and issues, especially in current era, lot of advancement has been taken place in terms of techniques and datasets use for detection and recognition of text.

The purpose of this survey is (1) introduce recent works, (2) identify modern algorithms and datasets, (3) identify different adversarial attacks on text recognition, (4) identify the different libraries supporting text detection and recognition task, (5) next generation applications and predict research guidelines for the future. Additionally, this manuscript offer in detail comparative analysis of recently used techniques and benchmark datasets. This study can offer a noble orientation for beginner researchers in the field of detection and recognition of scene text.

Keywords; *Adversarial attack, Computer vision, Text detection, Text recognition*

I. INTRODUCTION

Text is one of the most significant discoveries of humanity, which has played a vital responsibility in human life since from earliest times. The amusing and specific information embodied in text is very helpful in a broad variety of applications such as searching image, targeting geo location, interaction of human computer, navigation of robot and industrial automation. As a product of individual notion, in natural scenes text holds high level meaning. This feature allows a text present in natural images and videos to analyze and understand the subsequent surroundings. With the growing popularity of elegant phones and movable networking devices, text data present in images is obtain more easily and proficiently. Consequently detection and recognition of text in natural scenes turn out to be key and hot research topic in the field of computer vision and analysis of document. Mainly in current years, the society has witness a tremendous research work and extensive advancement in this field. Additionally, International Conference for Document Analysis and Recognition (ICDAR) initiates “Robust Reading” competition in 2003, and since then many techniques have been proposed to deeply progress the development of detection and recognition of scene text. [4].

The past is a mirror for the future. What we are short of today tells us about what we can expect tomorrow. Despite the success so far, task of detection and recognition of scene text are still confronted with several challenges. Fig.1 displays some challenging examples of scene text. Human

Published by: The Mattingley Publishing Co., Inc.

TEST Engineering & Management
January - February 2020
ISSN: 0193-4120 Page No. 16917 - 16930

have barely no difficulties localizing and recognizing text, current algorithms did not design and trained easily. They have not yet reached human-level performance. Fig.2 displays the percentage of complexity cause by each challenge address by various authors [11, 50, 63, 30] in their manuscript which affects the process of recognition from the natural scenes [5]. The key challenges in the field of locating and identification of scene text can be approximately classify into three types as text diversity, background complexity and Interference factors which are analyzed as follows

Text diversity: Scene text varies in color, size, orientation, font, language, and partial text deletion, etc.

Fig. 1 Samples of challenges in text understanding. a Different orientation; b Complexity of background; c. blur

d different colors; e. different size; f Different font style; g Light Intensity; h Ignorance of some text part; i. different languages.

Background complexity: The backgrounds in different scenery images and videos can be very compound. Items such as signs, boundary, grasses and blocks are nearly not separable from real text, so it causes confusions and errors.

Interference factors: Various obstruction factors, for instance, blur, distortion, very low resolution, noise non-uniform Illumination and partial occlusion may increase failure in scene text detection and recognition. [2, 4, 5]

To deal with challenges mentioned above, number of techniques has been proposed and in current year's significant progresses have been achieved [6].

The remaining manuscript is ordered as given below Section 2 offers review of previous detection and recognition of text techniques, while section 3 describes few significant techniques. Whereas section 4 represents various benchmark datasets used in literature, Section 5, discuss about various libraries and tools which can be used to support text detection and recognition task. Section 6 focuses on various adversarial attacks. Section 7 presents performance of existing techniques with respect to benchmark datasets, while Section 8 points out some potential research direction and finally section 9 concludes the paper.

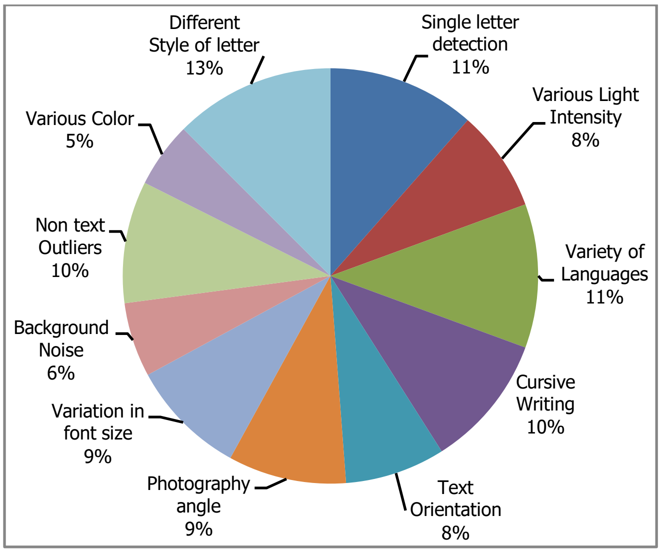

Fig. 2 Complexity of Challenges in Scene Text Understanding

II. RECENT DEVELOPMENT FOR DETECTION AND RECOGNITION OF TEXT

This task mainly involves three tasks, namely text detection, recognition of text and end to end spotting of text [2]. In this segment, we will argue and investigate the conventional and progressive techniques available for detecting and recognizing of text with respect to three tasks mentioned above. Furthermore, the techniques depend on deep learning described and evaluated in depth. Fig 3

January - February 2020
ISSN: 0193-4120 Page No. 16917 - 16930

shows classification of various techniques for detection & recognition of text.

A. Text Detection

This task plan to locate text in images and create bounding boxes for words. Previous techniques can be roughly classified into three categories: character based, text-line based and word based methods.

a. Sliding window-based technique

In Sliding window-based technique, text information is detected via moving a multi scaling sub window through all probable spot in an image, and then employ a proficient classifier to recognize whether text is present inside the sub window [3,10].

Babenko has proposed mechanism of an end to end pipeline, where they use SW classifier to carry out multi-scale character detection. Pan et al. has anticipated the text existing confidence and scale information through SW.

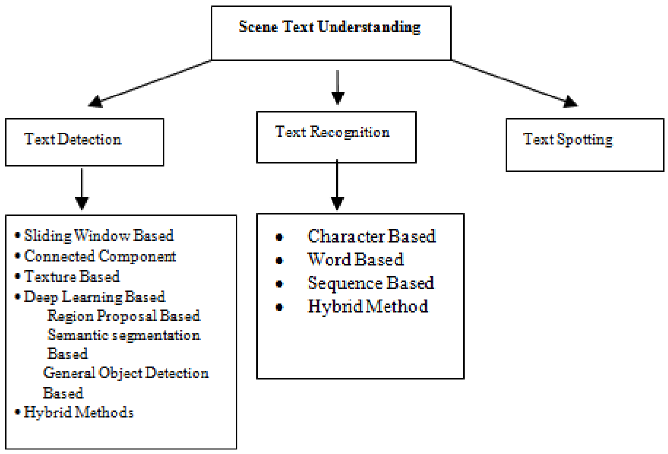

Fig. 3 Classification of Detection& Recognition of Text Techniques

Wang make use of convolutional neural network model through SW scheme to get candidate lines of text in certain image, and then guess text spot.

b. **Connected component (CC)-based technique**

These techniques finds and combines small constituents into a large constituent, after that sort out non text constituents through classifier, and lastly take out text from image and unite it into text region. These techniques have advantage of small calculation and efficiency. Though, they are afflicted by a number of cons as the inability to handle rotation, scale changes, complex backgrounds and other puzzling cases. With reference to CC-based methods, the most representative techniques are maximally stable extreme regions (MSERs) and stroke width transform (SWT) [2,14].

Epshtein presented stroke width transform SWT operator that make use of Canny edge detector to carry out edge detection & figure out per pixel width of utmost probable stroke containing pixel. Neumann et al. make use of maximally stable extremal regions (MSERs) algorithm to detect text. They first produce candidate region and then categorize it through the pre trained character and non-character classification. Lastly, the model produces the text line. Yin developed a fast MSERs pruning algorithm, that dig out character candidate and then AdaBoost classifier is used to detect text.

c. Texture-based technique

This technique treats texts as a distinct type of texture and makes use of their textural features, such as local intensities, filter responses and wavelet coefficients, to differentiate between textual and non-textual parts in the images. As all places and scales should be scanned, these techniques are usually computationally costly. Also these techniques mostly grip horizontal texts and are profound to rotation and scale change [4,19].

Zhong developed a filter which can openly detect text in the discrete cosine transform (DCT) domain. This algorithm runs fast but it has comparatively low detection correctness. Hanif has given a cascade scene text detector with neural network-based localizer to understand localization rules

Published by: The Mattingley Publishing Co., Inc.

TEST Engineering & Management

January - February 2020
ISSN: 0193-4120 Page No. 16917 - 16930

automatically. This technique is appropriate for detection of text with numerous styles and sizes localizer. Zhang has proposed a symmetry-based text line detector which has the ability to locate text lines from images.

d. Deep learning-based technique

Currently, deep learning has been utilized broadly in various techniques for general object detection and semantic segmentation and attained grand success. Recently CNN has been broadly explored as it is insensible to geometric transformation, deformation and illumination with small computational cost and it is widely used for text understanding due to its various properties. Most of existing CNN based techniques are classified into region proposal-based technique, segmentation-based technique, General object detection-based technique [23].

• Region proposal-based technique

It takes benefits of the region based techniques plus the deep neural networks. These techniques can strongly study a module representation through CNN in addition to it also make complete use of CNN's dominant capability in object detection and classification [22].

Huang introduced a strong text detector that united maximally stable extremal regions (MSERs) with convolutional neural network (CNN). It has achieved good result on ICDAR 2011. Zhang introduced a framework which can locate text in multiple orientations and languages. It broadly consists of two fully convolutional network (FCN): one of which is used for the salient map of text regions and the other FCN is used to forecast the centroid of each character.

•Semantic segmentation-based technique

The semantic segmentation-based detectors firstly dig out text blocks from the segmentation map which is created through (FCN). Afterward, bounding boxes of text are acquired by complex post-processing [26].

Qin et al introduced a cascaded technique with two convolutional neural networks. One is a fully convolutional network known as TextSegNet, it is designed to locate text blocks. And other is a YOLO-like network name as WordDetNet which locate individual words by producing oriented rectangular regions. Tang et al. proposed cascaded convolutional neural network (CNNs) where 3 CNN-based models were introduced as detection network (DNet), segmentation network (SNet) and classification network (CNet). DNet used to identify areas of text and CNet used to obtain true text region. Wang et al introduced deep convolutional neural networks (DCNN) classifier which has double threshold strategy to categorize text or non-text constituents. He et al. proposed cascaded convolutional text networks (CCTN) that make use of two networks to employ coarse-to-fine segmentation for scene image.

•General object detection-based technique

General object detectors forecast candidate bounding boxes directly by regarding texts as objects. Dissimilar from common objects, texts have clear definition of orientation that should be forecast in addition to the axis-aligned bounding box information [3].

Gupta et al. introduced proficient engine which could produce synthetic scene images with text annotations and all synthetic images were utilized to train a fully-convolutional regression network (FCRN) for localizing text. Tian presented a connectionist text proposal network (CTPN) to locate text. In CTPN, for feature extraction, VGG16 is first used, and then a vertical anchor mechanism is created to forecast localizing text in a fine scale. Liao et al. introduced an end-to-end trainable scene text detector known as TextBoxes that is motivated by SSD. Text-box layers were incorporated in the architecture of Text Boxes, which could locate the words with extreme aspect ratios by designing long default boxes and irregular 1*5 convolutional filters. Ma et al suggested a rotation region proposal

Published by: The Mattingley Publishing Co., Inc.

TEST Engineering & Management
January - February 2020
ISSN: 0193-4120 Page No. 16917 - 16930

networks (RRPN) that is assemble upon the Faster-RCNN architecture. Jiang et al. also produced a Faster RCNN based framework, known as rotational region CNN (R2CNN), for multi-oriented detection of text.

c. Hybrid technique

Currently, few researchers tried to merge the two kinds of above techniques as hybrid technique so as to correctly locate texts in more difficult situations.

Turki offered an efficient and robust scene text detector. They have applied Otsu technique and a robust edge projection to filter the complex background. MSER technique was used to locate a text pixels candidate. He proposed a innovative text attentional convolutional neural network (Text-CNN) that is successfully used for acquiring text-related regions and features from the image constituents. Fabrizio presented a context free technique for detection of text. Precisely, CC based method was utilized to create text candidates. Texture-based technique was employed to authenticate or reject the generated candidates. This technique is employed to multi-sized, multi-oriented texts. Zhong introduced an anchor-free region proposal network (AF-RPN) that could produce excellent inclined text proposals directly without designing complicated hand crafted anchors.

We have found that now a day’s most of authors are attracted towards the concept of deep learning and they have incorporated it in their proposed work.

B. Text Recognition

Text recognition mainly focused to identify the text in text candidate under the hypothesis that the text has been successfully sited. Text recognition approaches can be broadly classified into three categories: character-based techniques, word-based techniques and sequence-based techniques.

a. Character-based techniques

Character-based technique performs character level recognition of text. As the rudimentary element of text, character posses a huge number of vital information. Successful recognition of character builds bottom-up text recognition which is easy to implement [42].

Bissacco et al. make use of deep neural network which is trained on HOG features for character classification. So as to improve the recognition performance, a two-level language model is used. Jaderberg et al. introduced a CNN based framework which adopts a conditional random field (CRF) graphical model. Lee offered recursive recurrent neural networks (RNNs) with attention model for recognition of text. Yao suggested an advanced multi-scale representation for recognition of scene text, known as strokelets whose spirit is a set of multi-scale mid-level primitives.

b. Word-based techniques

Word-based technique identifies text at word level and attracts the care of many researchers. Few existing techniques treat recognition of word as an optimization problem [49,3].

Jaderberg offered CNN based framework that is trained by make use of synthetic data without handcrafted labeling and accomplish good performance for recognition of word. Kang et al. introduced a context-aware convolutional recurrent network for recognition of word. Weinman suggested a reading system where they have included segmentation of word with recognition in the probabilistic structure.

c. Sequence-based techniques

In recent times, many authors have considered the problem of recognizing text as a sequence recognition task, in which the text is denoted through character sequence. Sequence-based technique reflect the closely similarity between characters [52].

Published by: The Mattingley Publishing Co., Inc.

January - February 2020
ISSN: 0193-4120 Page No. 16917 - 16930

Shi presented a mechanism of convolutional recurrent neural network (CRNN) for image-based sequence recognition. Lee et al. suggested a lexicon-free photo-OCR system known as recursive recurrent neural networks with attention modeling (R2AM). Yang et al. mostly concentrated on irregular recognition of text task where they have offered an end-to-end model which includes an auxiliary dense character detection model and attention model.

d. Hybrid technique

Currently, few researchers tried to merge two kinds of above techniques as hybrid technique in order to properly recognize texts under more complex circumstances. [2]

Shi implemented word recognition with respect to a cost function which is well-defined by character detection scores, spatial constraints and linguistic knowledge. Almazan basically focused on word recognition technique that is based on embedded attributes. Lou suggested recognition of word model as a high-order factor graph with hypothetical neighboring candidate characters are constructed edges of the graph and taken as random variables. There are four factors as transition, smoothness, consistency, and singleton, are defined and used for word parsing.

We have found that Character classification-based techniques are not sensitive to noise, blur, font variation and orientation. Word classification can efficiently identify words in scene image with a number of class labels.

e. End-to-end text spotting

The end to end text spotting technique is an amalgamation of detection of text and recognition of text. In a unified system, the recognizer creates recognition outputs and also regularizes detection of text through its semantic-level awareness. A whole framework is applied to implement image input, text localization and recognition of text. [2]

Wang et al. projected CNN based model, which make use of NMS to eliminate overlapping candidates and acquire the set of line-level bounding boxes for texts. And then beam search technique is applied to discover the best segmentation of words. Jaderberg make use of two CNNs for bounding box regression and text recognition correspondingly. Moysset et al. suggested a CRNN based system, where the convolutional layers share parameters over the different repressors to discover text lines locally, and a 2D-LSTM model is skilled with CTC alignment to identify texts. Liao et al. anticipated an innovative text detector known as TextBoxes ++, that could proficiently detect multi-oriented scene text, it is united with a text recognizer, TextBoxes ++ also be applied for end to end spotting of text.

III. SIGNIFICANT TECHNIQUES FOR DETECTION AND RECOGNITION OF TEXT

A. Fully Convolutional Network (FCN)

For effective semantic segmentation, FCN is a powerful visual model that can produce hierarchies of features. Due to assistances of multi-scale learning and forecast obey to the nature of text; various existing techniques employ FCN as their major support for detection of text. Generally, a pixel-wise text and non-text salient map is acquired by make use of FCN that creates pixel-wise labeling or labeled region holding texts. Succeeding to that, candidate bounding boxes of text are produced. [70].

B. You only Look Once (YOLO)

YOLO is a single convolutional network which at the same time forecasts a number of bounding boxes and class probabilities for those boxes in one evaluation. YOLO frame object detection as a regression problem to spatially distinct bounding boxes and allied class probabilities and it processes images in real-time at 45 frames per second, YOLO is tremendously fast compared with R-CNN based system. On the other hand, it may attain poor precision while locating small size objects. Hence it

January - February 2020
ISSN: 0193-4120 Page No. 16917 - 16930

may not be directly valuable for detection of text. [68].

C. Regions with CNN (R-CNN)

R-CNN is an end-to-end architecture for use for object detection. For R-CNN, an image and multiple regions of interest (RoIs) are input whiles of tmax probabilities and per-class bounding-box regression offsets are the outputs. Faster R-CNN marks enhancement over Fast R-CNN, which aims to decrease the time spent on region proposals generation [55].

D. Single Shot Detector (SSD)

Using a single deep neural network SSD determine objects in images. It describe a set of default boxes for the output space of bounding boxes, and along with it forecast the shape offsets and the confidences for all object groups. In SSD, forecasts are merged from multiple feature maps with different resolutions. SSD could proficiently handle up with objects of different sizes when compared with YOLO. As SSD incorporate the benefits of YOLO, Fast R-CNN and Faster R-CNN, different techniques make use of SSD for detection of text via including precise alteration, such as designing default boxes with multi orientations.[67].

IV. BENCHMARK DATASETS

Despite the success in detection & recognition of Scene text, existing techniques are evaluated and tested on various datasets after being trained on them separately. Basically a new dataset representing numerous challenges would also offer extra momentum for this field. Generally the model is trained only on one dataset and subsequently tested on another. We have collected various datasets and summarize their features in Table.1, whereas Figure 4 represents the sample images containing text with different challenge from commonly used benchmark datasets. Every dataset come up with several challenging images which need to be explored more in future. Art dataset is not

Fig. 4 Sample Benchmark Dataset

V. COMPUTER VISION TOOLS/LIBRARIES

Tools for computer vision have progressed over the past few years, now a day's computer vision is also being offered as a service. However, in the present day the enhancement in hardware like GPUs, machine learning tools and frameworks formulate the computer vision as more powerful. Key cloud service providers such as Microsoft, Google and AWS have all joined the race, to become the developers' choice.

But always we have question in our mind that which tool should we choose for implementation of our idea in computer vision. To get answer of this question we have summarized the pros and cons of various tools that can be used for computer vision as given below.

A. OpenCV

An OpenCV contains essential techniques and algorithms to carry out different image and video processing tasks. It is also a multiplatform tool support python and C++. When working with massive data sets or very large images OpenCV gets a bit slow. It relies on CUDA for GPU processing and it doesn't have GPU support.

Published by: The Mattingley Publishing Co., Inc.

TEST Engineering & Management

January - February 2020
ISSN: 0193-4120 Page No. 16917 - 16930

B. Matlab:

As compared to C++, it allows quick prototyping & its code is quite brief, making it easier to read and debug. It deals with errors before execution by suggesting few ways to make the code faster, but it is a paid tool.

C. TensorFlow:

It assists us in fetching the power of Deep Learning to computer vision and it has few special tools to carry out image processing classification. But it is very resources hungry and can consume a GPU's capabilities in no time.

D. CUDA

It is a platform used for parallel computing. It facilitates in computing performance by using the power of GPUs. The CUDA Toolkit consists of the NVIDIA performance primitive library which is a gathering of signal, image, and video processing functions. We can prefer to use CUDA, If we have large images to process that are GPU intensive, but it is tremendously high on power consumption.

E. SimpleCV

It allows us to access a large number of computer vision tools similar to OpenCV. If we want to get our work done easily without going into the depths of image processing, this is the tool can be used but community forum is not that much active.

F. GPUImage

Its an iOS library which permits us to apply GPU-accelerated effects and filters to images, movies and live motion video.

Table 1. Benchmark Dataset

Data Set	Images Training / Testing	Orientation	Language
Art [85]	5603/4563	Multioriented	English, Chinese
LSVT [77]	30K/20K	Multioriented	Chinese
SVT(2010) [44]	100/250	Horizontal	English
ICDAR 2003 [41]	258/ 251	Horizontal	English
ICDAR 2005 [3]	1001/489	Horizontal	English
ICDAR 2011 [2]	229/255	Horizontal	English
ICDAR 2013 [38]	229/233	Horizontal	English
ICDAR 2015 [64]	1000/500	Multioriented	English
RCTW(2017) [70]	8034/4229	Multioriented	Chinese
Total-Text (2017) [5]	1255/300	Curved	English, Chinese
CTW (2017) [6]	25K/6K	Multioriented	Chinese
IIIT 5k Word [84]	2000/3000	Horizontal	English
COCO-TEXT (2017) [59]	63686/43686	Multioriented	English
SVHN [67]	73257/26032	Horizontal	English
MSRA-TD500 [28]	300/200	Multioriented	English, Chinese

VI. ADVERSARIAL ATTACK

Most of the existing machines learning algorithms use for classification task are greatly susceptible to adversarial instance. In an adversarial instance, an input data has been altered to some extent in such a way that the machine learning classifier misclassifies it. In majority of cases, these alterations are so small that a general observer does not even note the alteration at all. Adversarial instance causes security concern since they could be utilized to accomplish an attack on systems with machine learning, though the adversary has no right to access the specific model. To carry out investigation on adversarial attacks plus defenses against it, is not easy for several causes. One cause is that evaluation of suggested attacks and defenses is not easier. Different adversarial attacks can be classified along various dimensions as given below

A. Based on type of outcome the adversary desires:

Non-targeted attack: In this case intruder's objective is to cause the classifier to forecast any wrong label.

Targeted attack: In this case intruder intends to alter the classifier's forecast to specific target class.

Published by: The Mattingley Publishing Co., Inc.

January - February 2020
ISSN: 0193-4120 Page No. 16917 - 16930

B. Based on adversary's amount of knowledge about the model:

White box: In this case intruder has complete information about the model with model architecture, model type and values of all parameters and trainable weights.

Black box with probing: In this case, adversary doesn't have in depth knowledge concerning the model, but can query the model. Adversary can feed a few inputs and observe outputs.

Black box without probing: In this case, the adversary has no information related to model under dataset. But can attack by constructing adversarial examples which can deceive the majority of machine learning models.

Based on way of data feeding into the model by adversary:

Digital attack: Here the intruder can directly access to the actual data fed into the model.

Physical attack: In this case the model fed an input acquire by sensors such as microphone or camera where adversary can change the exact digital representation acquired by the sensors based on aspect identical to camera angle, ambient light or sound in the surroundings. [7].

VII. PERFORMANCE EVALUATIONS AND COMPARISON

In this section, we discuss about various performance metrics used to assess performance of various detection and recognition of scene text techniques such as Precision, Recall, F-Measure and accuracy and finally compared the result obtain by existing proposed techniques.

A. Result Comparison

The result obtained by existing techniques for detecting and recognizing of scene text with respect to various commonly used benchmark datasets such as ICDAR2003, ICDAR 2013, ICDAR 2015, COCO Text, SVT, MSRA-TD500 are listed and compared in Tables 2 and 3. The P, R, F mention in the table 2 and 3 denote precision, recall and F measure, respectively. "50" is lexicon size, whereas "Full" indicate the joint lexicon of all images in the benchmarks, and "None" denote lexicon-free.

For text detection task, done by various authors we have found that Zhong et.al obtained best result with ICDAR2013 dataset, Liao et.al got best result for COCO Text dataset, Yang et.al obtained best result with ICDAR 2015 and Lyuet. al got best result for MSRA-TD500 dataset as compared to result obtain by other authors with same dataset. Similarly for text recognition task we have found that Jaderberg et.al obtained best result for ICDAR2003 dataset and Baiet.al got best result with ICDAR2013, ICDAR2015 and SVT dataset as compared to result obtain by other authors with same dataset. Results listed in bold in Table 2 and Table 3 represent best result.

VIII. RESEARCH DIRECTIONS

In recent year's considerable progress have been done in detection and recognition of text. Still there are number of challenges that should be considered in the future. There is still a long way to go to create practical systems which can correctly and vigorously extract textual information from natural scenes as that of human being; we consider the following facets are precious to investigate in the next decade.

A. Multi-Orientation:

In real-world circumstances, texts can be in diverse orientations. However many researchers in this fields only consider their interest to horizontal texts. In order to obtain full benefit of textual information in natural scenes, it is essential to read texts of different orientations [2].

B. Multi-Language:

Detecting and recognizing of scene text in surroundings is a key factor for several applications,

Published by: The Mattingley Publishing Co., Inc.

TEST Engineering & Management
January - February 2020
ISSN: 0193-4120 Page No. 16917 - 16930

ranging from shop indexation in a street to business card digitization. This new competition intends to review the capability of high-tech technique to detect and recognize multi-lingual text. Most of the existing techniques concentrated on text in English, it is vital to build up such recognition systems which can deal with texts of diverse languages [23].

C. Deep Learning & Big Data:

More enhancements in terms of accuracy can be achieved in the field of detection and recognition of text, if we use deep learning structure to determine and model the features of scene text from huge size of data [6].

D. Ignorance of some Text Part

Ignoring certain chunks in object detection may be adequate as it rests semantically the same, but then again it would be terrible for the final text recognition results as few characters may be missing, resulting in different words. A corner of a building without any difficulty recognized as I or wheels of any vehicle as O. This challenges the text detection techniques to differentiate text from non-text.

E. Visual Question Answering

It stresses on a precise type of Visual Question Answering task, where understanding the text information in a scene is essential in order to give an answer.

F. Towards Stable Performance: as Needed in Security

Recent study have shown that deep learning models are vulnerable to adversarial attack so we have to pay more consideration to possible security risks, particularly, when applied in security services such as identity check [6].

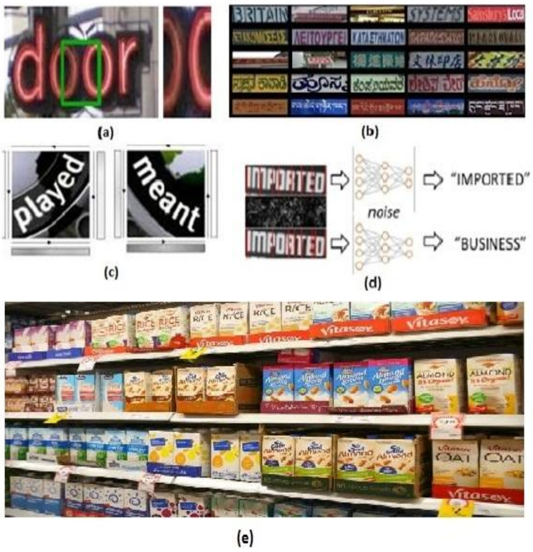

Fig. 5 Research directions.a ignorance of some text b Multilanguage text c multi oriented text d. Adversarial attack e. Visual Question Answering

Table 2Text Detection Result

Sr No	Technique	Year	ICDAR 2013			ICDAR 2015			COCO Text			MSRA-TD500		
			P(%)	R(%)	F(%)	P(%)	R(%)	F(%)	P(%)	R(%)	F(%)	P(%)	R(%)	F(%)
1	Deng [64]	2018	88.6	87.5	88.1	85.5	82	83.7				83	73.2	77.8
2	Liao [32]	2018	92	86	89	87.8	78.5	82.9	**64**	**57**	**61**	87	73	79
3	Li [7]	2018				89.3	85.2	87.2						
4	Zhong [16]	2018	**94**	**90**	**92**	89	83	86						
5	Lyu [35]	2018	92	84.4	88	89.5	79.7	84.3	61.9	32.4	42.5	**87.6**	**76.2**	**81.5**
6	He [33]	2018	91	89	90	87	86	87						
7	Yang [42]	2018				**93.8**	**87.3**	**90.5**				87.5	79	83
8	Liao [32]	2017	88	82	85	88	83	85						
9	Zhou [59]	2017				83.3	78.3	80.7	50.4	32.4	39.45	87.28	67.4	76.08
10	He [33]	2017	89	86	88	80	73	77	46	31	37			
11	Ma [30]	2017	90	72	80	82.2	73.2	77.4				82.1	67.7	74.2
12	Shi [39]	2017	87.7	83	85.3	73.1	76.8	75				86	70	77
13	Yao [6]	2016	88.88	80.2	84.3	72.3	58.7	64.7	43.2	27.1	33.31	76.51	75.3	75.91

Published by: The Mattingley Publishing Co., Inc.

January - February 2020
ISSN: 0193-4120 Page No. 16917 - 16930

14	Tian [28]	2016	93	83	88	74	52	61						

Table 3Text Recognition Result

Sr No	Technique	Year	ICDAR 2003			ICDAR 2013	ICDAR 2015	SVT
			50	Full	None	None	None	None
1	Bai [04]	2018	98.7	97.9	94.6	**94.4**	**73.9**	**87.5**
2	Yang [42]	2017				85.21	79.78	
3	Lou [48]	2016				86.2		80.7
4	Jaderberg [41]	2016	**98.7**	**98.6**	**93.3**	90.8		80.7
5	Lee [46]	2016	97.9	97	88.7	90		80.7
6	Shi [61]	2016	98.3	96.2	90.1	88.6		81.9
7	He [51]	2016	97	93.8				
8	Shi [50]	2015	98.7	97.6	89.4			80.8
9	Yao [22]	2014	88.5	80.33				75.89
10	Shi [57]	2013	87.4	79.3				73.51
11	Wang [49]	2012	90	84				

IX. CONCLUSION

Textual content in human environments carries vital information that is unavailable in any other form in the scene. Understanding written information in human surrounding is crucial to carry out our day to day tasks like purchase of item and using public transport.

Detection and recognition of scene text have acknowledged increasing attention due to its numerous applications in the arena of computer vision based system. Past is a mirror for the tomorrow, what we be deficient in today tells us about what we can be expecting tomorrow. This paper primarily reviewed and classified various techniques proposed in last decade for detection and recognition of text, also key techniques are highlighted. We have recognized that deep learning has resulted in great success in terms of the performance with respect to reference datasets. In spite of the victory so far, techniques for detection and recognition of text are still face up to numerous challenges. They have not yet reached up to human level performance.

We have analyzed and compared the performances of widely used techniques. Several widely used benchmark datasets are analyzed & results obtained by various techniques with respect to benchmark datasets are compared. Compared with previous surveys, we have analyzed and classified various potential security risks term as adversarial attack, also we have analyzed various libraries use for detection & recognition of text and finally we discussed different future trends in this field of detection and recognition of text.

REFERENCES

1. Neumann L., Matas J., “A method for text localization and recognition in real-world images”, Asian Conference on Computer Vision, 770–783. Springer, 2010.
2. Liu X, Meng G, Pan C, “Scene text detection and recognition with advances in deep learning:a survey”, International Journal on Document Analysis and Recognition (IJDAR) 2019.

3. Lin H, Yang P, Zhang F," Review of Scene Text Detection and Recognition", Archives of Computational Methods in engineering 2019.
4. Zhu Y, Yao C, Bai X, "Scene Text Detection and Recognition, Recent Advances and Future Trends", Article in Frontiers of Computer Science (print) 2016.
5. Francis L, Sreenath N, "Robust scene text recognition using manifold regularized Twin-Support Vector Machine", Journal of King Saud University –Computer and Information Sciences 2019.
6. Long S, He X, Yao C, "Scene Text Detection and Recognition: The Deep Learning Era", Draft Version Submitted to Cornell University 2018.
7. Yuan X, He P, Li X, "Adaptive adversarial attack on scene text recognition", arXiv preprint arXiv:1807.03326 2018.
8. Pan YF, Hou X, Liu CL, "A hybrid approach to detect and localize texts in natural scene images", 20:800–813 2011.
9. Mishra A, Alahari K, Jawahar CV, "Scene text recognition using higher order language priors", In: Proceedings british machine vision conference, pp 1–11 2012.
10. Wang T, Wu DJ, Coates A, Ng AY, "End-to-end text recognition with convolutional neural networks", In: International conference on pattern recognition, pp 3304–3308, 2012.
11. Jaderberg M, Vedaldi A, ZissermanA , " Deep features for text spotting", In: European conference on computer vision, pp512–528, 2014.
12. Epshtein B., OfekE,Wexler Y, "Detecting text in natural scenes with stroke width transform", In IEEE Conference on CVPR, pp. 2963–2970, 2010.
13. Yin X, Huang K, Hao H, "Robust text detection in natural scene images", IEEE Trans. PAMI 36(5), 970–983, 2014
14. Neumann L, Matas J, "Real-time scene text localization and recognition", IEEE Conference on Computer Vision & pattern recognition, 2012.
15. Cho H, Sung M., Jun B," Canny text detector: fast and robust scene text localization algorithm", Conference on Computer Vision & pattern recognition, 2016.
16. Zhong Y, Zhang H, Jain K,"Automatic caption localization in compressed video", IEEE Trans.action PAMI 22(4), 385–392, 2000.
17. Hanif S, Prevost L, Negri P," A cascade detector for text detection in natural scene images", In: ICPR, pp. 1–4, 2008.
18. Zhang Z, Yao C, Bai X., "Symmetry-based text line detection in natural scenes", Conference on Computer Vision & pattern recognition, pp. 2558–2567, 2015.
19. Liang G., Shivakumara P, Lu T, Tan L., "A new wavelet laplacian method for arbitrarily-oriented character segmentation in video text lines", In: ICDAR'15, pp. 926–930. IEEE, 2015.
20. Chen X, Yuille A, "Detecting and reading text in natural scenes", In Proceeding of Conference on Computer Vision & pattern recognition, 2004
21. Huang W, Qiao Y, Tang X," Robust scene text detection with convolution neural network induced mser trees", In: ECCV, pp.497–511.Springer, 2014.
22. Zhang Z, Zhang C, Shen W, Yao C, Liu W, Bai X," Multioriented text detection with fully convolutional networks", Conference on Computer Vision & pattern recognition, pp. 4159–4167, 2016.
23. Qin S, Manduchi R., "Cascaded segmentation-detection networks for word-level text spotting", arXiv preprint arXiv:1704.00834, 2017.
24. Tang Y, Wu X., "Scene text detection and segmentation based on cascaded convolution neural networks", IEEE Transaction on. Image Processing.26(3), 1509–1520, 2017.
25. Wang C, Liu C," Scene text detection with novel superpixel based character candidate extraction", In: ICDAR'17, vol. 1,pp. 929–934, 2017.
26. Dai Y, Huang Z, Gao Y, Xu Y, Chen K," Fused text segmentation networks for multi-oriented scene text detection", arXiv :1709.03272, 2017
27. Gupta A, Vedaldi A, Zisserman A," Synthetic data for text localization in natural images", In: IEEE conference on computer vision and pattern recognition, pp 2315–2324, 2016.
28. Tian Z, Huang W, He T, He P, Qiao Y," Detecting text in natural image with connectionist text proposal network", In: European conference on computer vision, pp 56–72 2016.

29. Liao M, Shi B, Bai X, Wang X, Liu W,” TextBoxes a fast text detector with a single deep neural network”, In: Proceeding of association for the advancement of artificial intelligence, pp 1–7, 2017.
30. Ma J, Shao W, Ye H, Wang L, Wang H,” Arbitrary-oriented scene text detection via rotation proposals”,. IEEE Transaction on Multimedia, 2017.
31. Jiang Y, Zhu X, Wang X, Yang S, Li W: R2CNN,” rotational region cnn for orientation robust scene text detection”, arXiv :1706.09579, 2017
32. Liao M, Zhu Z, Shi B, Xia G, Bai X,” Rotation-sensitive regression for oriented scene text detection”, IEEE conference on computer vision and pattern recognition, pp 1–10, 2018.
33. He T, Huang W, Qiao Y, Yao J,” Text-attentional convolutional neural network for scene text detection”, IEEE Transactionon Image Processing. 25(6), 2529–2541, 2016.
34. Zhong Z, Sun L, Huo Q,” An Anchor-Free Region proposal network for faster R-CNN based text detection approaches”, arXiv :1804.09003, 2018.
35. Lyu P, Yao C, Wu W, Yan S, Bai X,” Multi-oriented scene text detection via corner localization and region segmentation”, IEEE conference on computer vision and pattern recognition, pp 1–10, 2018.
36. Bissacco A, Cummins M, Netzer Y, Neven H,” Photo OCR reading text in uncontrolled conditions”, IEEE international conference on computer vision, pp 785–792 2013.
37. Jaderberg M, Simonyan K, Vedaldi A, Zisserman A,” Deep structured output learning for unconstrained text recognition”, In International conference on learning representations, pp 1–10, 2015.
38. Lee CY, Osindero S,” Recursive recurrent nets with attention modeling for OCR in the Wild”, IEEE conference on computer vision and pattern recognition, pp 2231–2239, 2016.
39. Yao C, Bai X., Shi B, Liu W: Strokelets,” A learned multiscale representation for scene text recognition”, In: CVPR, pp. 4042–4049, 2014.
40. Lou X, Kansky K, Lehrach W, Laan C, Marthi B, Phoenix D, George D,” Generative shape models: joint text recognition and segmentation with very little training data”, In: NIPS, pp. 2793–2801, 2016.
41. Jaderberg M, Simonyan K, Zisserman A,” Synthetic data and artificial neural networks for natural scene text recognition”, In: Conference on neural information processing systems, 2014.
42. Yang C, Yin XC, Li Z, Wu J, Guo C,” AdaDNNs: adaptive ensemble of deep neural networks for scene text recognition”, arXiv :1710.03425, 2017.
43. Phan, T, Shivakumara, P, Tan C,” Recognizing text with perspective distortion in natural scenes”, In: ICCV, pp. 569–576, 2013.
44. Weinman, J.J., Butler, Z., Knoll, D., Feild, J., “Toward integrated scene text reading”, IEEE Transaction. PAMI 36(2), 375–387, 2014.
45. He P, Huang W, Qiao Y, Loy CC, Tang X,” Reading scene text in deep convolutional sequences”, Proceedings of association for the advancement of artificial intelligence, pp 1–8, 2016.
46. Lee Y, Osindero S.,” Recursive recurrent nets with attention modeling for OCR in the wild”, In: CVPR, pp. 2231–2239, 2016.
47. Almazan J, Gordo A, Fornes A, ValvenyE,” Word spotting and recognition with embedded attributes”, IEEE Transaction on Pattern Analysis Machine Intelligence 36:2552–2566, 2015.
48. Lou X, Kansky K, Lehrach W, Laan V, Marthi B,” Generative shape models joint text recognition and segmentation with very little training data”, Advances in Neural Information Processing Systems. Neural Information Processing Systems, 2016.
49. Shi C, Wang C, Xiao B, Zhang Y, Gao S”, Scene text recognition using part-based tree-structured character detection,” IEEE conference on computer vision and pattern recognition, pp 2961–2968, 2013.
50. Jaderberg M, Simonyan K, Vedaldi A, Zisserman A,” Reading text in the wild with convolutional neural networks”, International Journal Computer Vision 116:1–20, 2016.
51. Liao M, Bai X, “TextBoxes ++ a single-shot oriented scene text detector”, IEEE Transaction Image Processing 27:3676–3690, 2018.

TEST Engineering & Management
January - February 2020
ISSN: 0193-4120 Page No. 16917 - 16930

52. B Moysset, C Kermorvant, C Wolf," Full-Page Text Recognition Learning Where to Start and When to Stop", International Conference on Document Analysis and Recognition, pp 871-876, 2017.
53. Liu X, Liang D, Yan S, Chen D, Qiao Y, "FOTS Fast oriented text spotting with a unified network", pp 1-10. arXiv :1801.01671, 2018.
54. Neumann, L., Matas, J.," Scene text localization and recognition with oriented stroke detection",. In: ICCV, pp. 97–104, 2013.
55. Yao, C., Bai, X., Liu, W,"A unified framework for multioriented text detection and recognition", IEEE Transaction on Image Processing.23(11), 4737–4749, 2014.
56. Li X, Wang W, Hou W, Liu RZ, Lu T," Shape robust text detection with progressive scale expansion network", pp 1–12.arXiv :1806.02559, 2018.
57. Yang Q, Cheng M, Zhou W, Chen Y, Qiu M," IncepText: a new inception-text module with deformable pooling for multi-oriented scene text detection", In: International joint conference on artificial intelligence, pp 1–7, 2018
58. Qin S, Manduchi R," Cascaded segmentation-detection networks for word-level text spotting", International conference on document analysis and recognition, pp 1275–1282, 2017.
59. Zhou X, Yao C, Wen H, Wang Y, Zhou S," EAST an efficient and accurate scene text detector", IEEE conference on computer vision and pattern recognition, pp 2642–2651, 2017.
60. He W, Zhang XY, Yin F, Liu CL," Deep direct regression for multi-oriented scene text detection", IEEE international conference on computer vision, pp 745–753, 2017.
61. Shi B, Bai X, Belongie S," Detecting oriented text in natural images by linking segments", IEEE conference on computer vision and pattern recognition, pp 2482–3490, 2017.
62. Yao C, Bai X, Sang N, Zhou X, Zhou S," SceneText detection via holistic, multi-channel prediction",. arXiv :1606.09002 pp 1–10, 2016.
63. Zhang Z, Zhang C, Shen W, Yao C, Liu," Multioriented text detection with fully convolutional networks", Computer vision & pattern recognition, pp 4159–4167, 2016
64. Bai F, Cheng Z, Niu Y, Pu S, Zhou S," Edit probability for scene text recognition", pp 1–9. arXiv :1805.03384 2018
65. Yao C, Liu W," Strokelets a learned multi-scale representation for scene text recognition", In: IEEE conference on computer vision and pattern recognition, pp 4042–4049, 2014.
66. Furkan A, Tito R, Mafla A, Gomez L, Rusiñol M, Mathew M, Jawahar C, Karatzas D," ICDAR 2019 Competition on Scene Text Visual Question Answering,, arXiv:1907.00490, 2019.
67. Long J, Shelhamer E, Darrell T," Fully convolutional networks for semantic segmentation", In: IEEE conference on computer vision and pattern recognition, pp 3431–3440, 2015.
68. He K, Zhang X, Ren S, Sun J," Deep residual learning for image recognition", pp 1–12. arXiv :1512.03385, 2015.
69. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S SSD," single shot multibox detector", In: European conference on computer vision, pp 21–37, 2016.
70. Girshick R," Fast R-CNN", In IEEE international conference on computer vision, pp 1440–1448, 2015.
71. Ren S, He K, Girshick R, Sun J," Faster R-CNN: towards real-time object detection with region proposal networks", In: International Conference on Neural Information Processing Systems, pp 91–99 9, 2015.
72. ChngC,LiuY,SunY,NgC,Fang C, ZhangS, Han J, Ding E, Chan C, Jin L," ICDAR 2019 Robust Reading Challenge on Arbitrary-Shaped Text-RRC-ArT", ICDAR 2019 to appear, 2019.