Scene Text Recognition: A Review

Article *in* Iraqi Journal of Science · January 2023

DOI: 10.24996/ijs.2023.64.1.37

2 authors, including:

Nidhal Khdhair El abbadi
Al Mustaqbal University
112 PUBLICATIONS **602** CITATIONS

ISSN: 0067-2904

Scene Text Recognition: A Review

Essam Haider Mageed[1]*, Nidhal Khdhair El Abbadi[2]
[1]*Department of computer science, college of computer science and mathematics, university of kufa , Najaf , Iraq*
[2]*Department of computer science, Collage of education, university of kufa, Najaf, Iraq*

Received: 14/10/2021 Accepted: 9/5/2022 Published: 30/1/2023

Abstract

The problem of text recognition and its applicability as part of images captured in the wild has gained a significant attention from the computer vision community in recent years. In contrast to the recognition of printed documents, scene text recognition is a difficult problem. Contrary to recognition of printed documents, recognizing a scene text is a challenging problem. Many researches focus on the problem of recognizing text extracted from natural scene images. Significant attempts have been made to address this problem in recent past. However, many of these attempts work on utilizing availability of strong context, which naturally limits the dictionary. This paper presents a review of recent papers related to scene text recognition in the period (2013-2020). This paper helps other researchers to understand the whole system of scene text recognition instead of reading many papers in isolation.

Keywords: text detection, text localization, text enhancement, segmentation, OCR.

التعرف على نص المشهد: مراجعة

عصام حيدر مجيد[1*], نضال خضير العبادي[2]
[1]قسم علوم الحاسوب, كلية علوم الحاسوب والرياضيات , جامعة الكوفة, النجف , العراق
[2]قسم علوم الحاسوب, كلية التربية, جامعة الكوفة , النجف , العراق

الخلاصة

اكتسبت مشكلة التعرف على النص قابلية تطبيقه كجزء من الصور الملتقطة في البرية اهتمامًا كبيرًا من مجتمع رؤية الكمبيوتر في السنوات الأخيرة. على عكس التعرف على المستندات المطبوعة ، يعد التعرف على نص المشهد مشكلة صعبة. التركيز على مشكلة التعرف على النص المستخرج من صور المناظر الطبيعية. بذلت محاولات كبيرة لمعالجة هذه المشكلة في الماضي القريب. ومع ذلك ، فإن العديد من هذه الأعمال تستفيد من توفر سياق قوي ، مما يحد بشكل طبيعي من القاموس. قدمت هذه الورقة مراجعة للورقة الأخيرة المتعلقة بالتعرف على نص المشهد في النطاق (2013−2020). تساعد هذه الورقة الباحث الآخر على فهم النظام الكامل للتعرف على نص المشهد بدلاً من قراءة العديد من الأوراق بشكل منفصل.

*Email: essamalmosawee@gmail.com

1. Introduction

Detecting text in natural images and business cards compared to scanning of printed pages, is considered as an important step for several computer vision applications, such as computerized assistance for the visually impaired, automatic corporate geocoding, and robotic navigation in urban environments. Text retrieval in both indoor and outdoor environments provides contextual clues to a variety of vision tasks. Moreover, it has been shown that performance of image retrieval algorithms critically depends on performance of their text detection modules[1].

Computer vision has a wide range of practical applications including autonomous driving, robotics and drones, mobile e-commerce, and assisting visually impaired. Image features play an important role in scene text recognition[2]. Many applications are very useful for scene text recognition. A very important one is mobile applications that are connected with analysis of media, content retrieval, scene content understanding, and assistant navigations system[3]. Another application is unnamed (vehicle-robot) navigation, living aids for visually impaired persons, and content-based image retrieval[4]. There are real-world applications that consist of street sign reading in a driverless vehicle, human-computer interactions, assistive technologies for blind and guide board recognition.

2. Challenges

The main problem of recognizing text in a scene is how to detect it inside an image. To do this, many challenges that must be taken into consideration to deal with all situation of scene text image. Figure 1 shows those Challenges. In general, challenges can be :

•Color: Characters of text may be all have the same color and this is easy to deal with it. On the other hand, characters with different colors are difficult to deal with.

•Character size: Size of a character is a very important because there are some situations where size changing may lead to unpredictability to recognize text.

•Font type: Many font types can be used; each one is working separately from others.

•Direction: It is very important because not all text is in the same direction.

•Overlap with background: When a text is overlapped with background, it is difficult to be detected.

•Huge background: Where content of a background is larger than content of another object in the same image.

•Big fluctuation in text appearance and scale: Some of the text may be scaled or rotated, this creates a problem in detection because there is no suitable text[5].

•Inter-character and intra character confusions: Means there is an interference between inter-characters of a word and other characters of another word[6].

•Multiple text forms: there are many situations for text appearance:

1.Horizontal

2.Oriented

3.Curved[7].

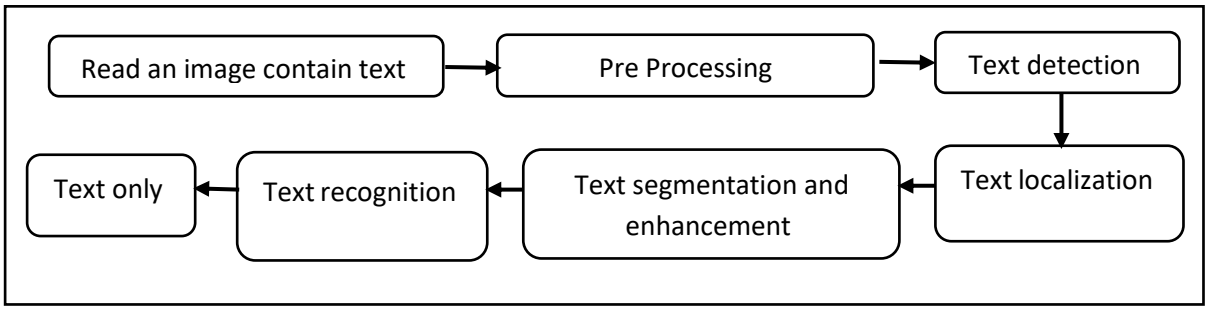
•Unusual font: this is very difficult to deal with because there is no fixed condition.

•Noise: many characters include noise and this is not easy to detect.

Figure 1: Text challenges from(a-j)[7]

3. Scene text recognition system

The steps of the scene text recognition consist of many important steps to do the task of text recognition. Figure-2 shows that. Take an Maximally Stable Extremal Regions (MSER) features to detect text and recognize the full processes show in Figure-3 [8-11].



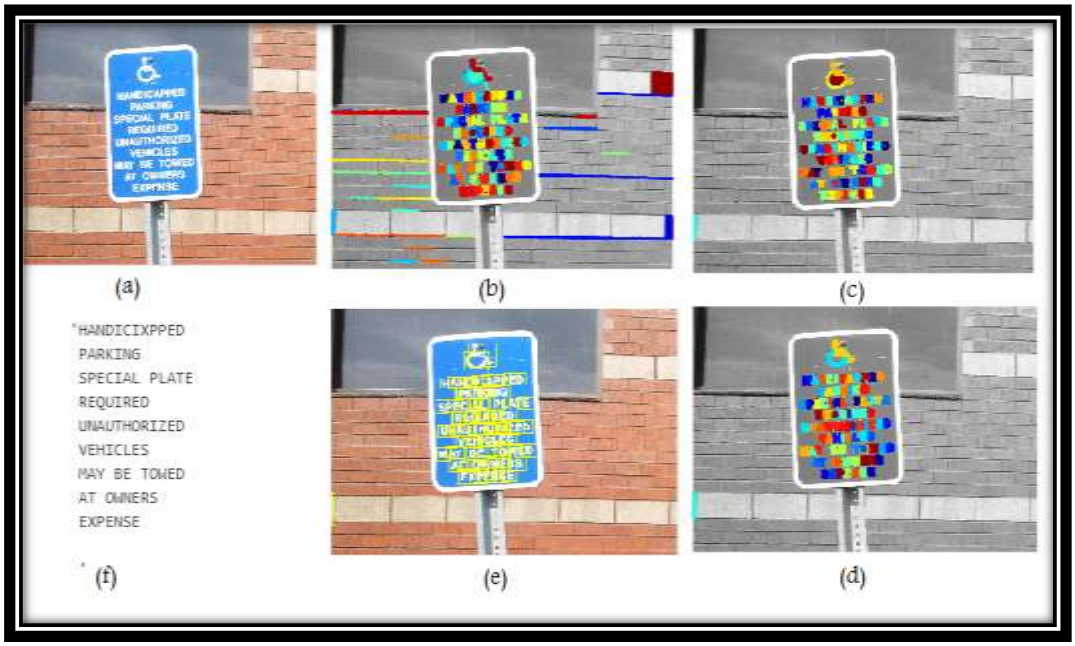


Figure 3: An example of the complete process of scene text recognition. a-original image. b-MSER regions. C- after removing non-text regions based on geometric properties. d- after removing non-text regions based on Stroke width variation. e-expanded bounding boxes text. f- recognize text by using OCR recognizer[8-11].

3.1 Pre-processing

Preprocessing step is necessary to obtain a better text recognition rate, using efficient algorithms of preprocessing creates the text recognition method robust using noise removal, image-enhancing process, image threshold process, skewing correction, page and text segmentation, text normalization, and morphological operations. The majority of OCR application uses binary/gray images. The images may have watermarks and/or non-uniform backgrounds that make recognition process difficult without performing the preprocessing stage. There are several steps needed to achieve this. The initial step is to adjust the contrast or to eliminate the noise from the image which is called image enhancement technique. The next step is to do thresholding to remove watermarks and/or noise followed by page segmentation for isolating graphics from the text. The next step is text segmentation to individual character separation followed by morphological processing. The morphological processing is required to add pixels if the preprocessed image has eroded parts in the characters [12]. Image mapping to distinct color channels red, green, and blue. Smoothing, noise elimination, and setting the image contrast [13]. The block diagram in Figure 4 shows that.

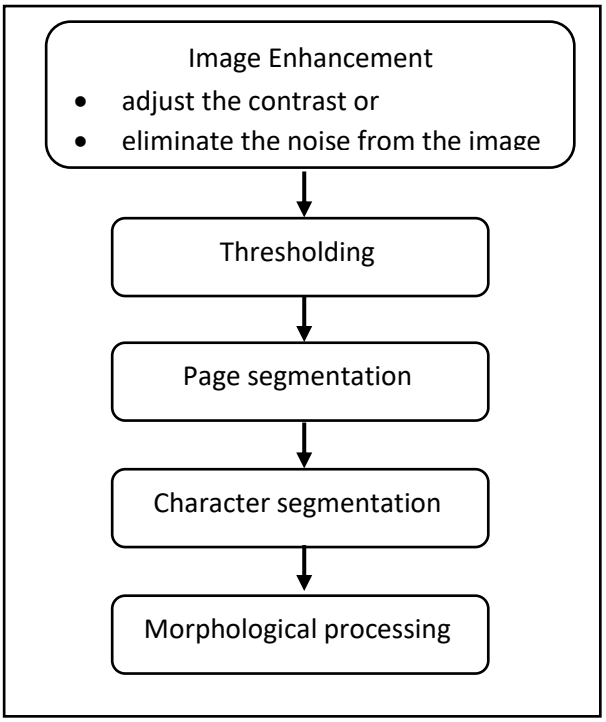


Figure 4: The Pre Processing parts

3.2 Text detection

This is the main step after preprocessing in scene text recognition. This means detecting and identifying text inside the image. Localization may be used in this step, but not always. Localization can do text detection in an image.

3.3 Text localization

This process finds locations of the detected text, and builds rectangles around them.

3.4 Text verification

During the text localization, there are many characters or words detected as false. So the result of the previous stage may not contain any text. These locations of false detection is ignored. Verification is only for the true locations.

3.5 Text segmentation and enhancement

Text segmentation is the process of isolating text from background. The segmentation process is done based on the true location that was produced from verification stage. The result of segmentation needs to be enhanced by using some image enhancement techniques.

3.6 Text recognition

This stage is converting the result from previous steps into a real text. Generating the text based on the segmentation.

3.7 Processing the results

To get the word-level bounding boxes of original image, box candidates in the second stage should be furtherly processed. There are three steps in post-processing: first Map bounding box candidates in text region proposals back to the original image. Second Remove bounding boxes that cannot contain the entire text words. Third After mapping the valid bounding boxes back to the original image. We used Non-Maximum Suppression (NMS) to

remove the redundant predictions of word-level annotation languages like Latin [14]. Figure 5 shows that.

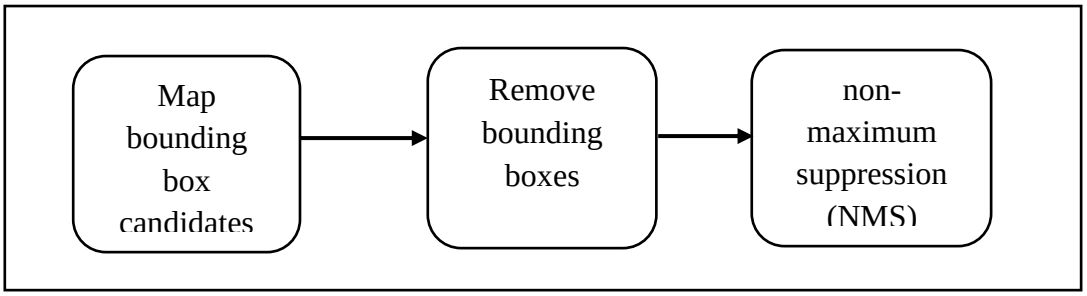


Figure 5: Post Processing parts

4. Literatures reviews

There are many papers published in the field of text detection and recognition. In this section, we introduce some of those papers, that focus on detection, localization, verification (as needed), enhancements, segmentation, and finally recognition (basically all methods depend on Optical Character Recognition (OCR)).

Wenhao He, Xu-Yao Zhang, Fei Yin, Zhenbo Luo , Jean-Marc Ogier , Cheng-Lin Liu[14]**.**

The proposed method is to take an image with text and apply scale based region proposal network (SPRN). This network is used to find text regions and then cuts it. There is something called scale estimation that has three levels (small, normal, and big) and the other thing is boundary regression **(as central stages).** After that, there is a proposal generation consist of four major things (split text region, box generation, box grouping, and box refinement). The next step is using a Fully convolutional layer (FCN) to localize the text and then map text bounding. The next step is post-processing consist of Mapping each bounding box, removing bounding boxes that cannot contain the entire text words, and Removing the predictions. Note the above-proposed method: the three scale estimation is very useful: small scale is used to detect the small text, big scale is used to detect big text, and using normal text to detect text between a small and big.

Hongwei Zhang a ,Changsong Liu ,Cheng Yang ,Xiaoqing Ding ,KongQiao Wang[15]**.**

The proposed method is based on two parts, the first one is based on the connected component (CC) with strict neighbor relational graph (S-NRG), and the second one is based on relaxed neighbor relational graph(R-NRG). The purpose of NRG is to isolate text regions from non-text regions. In this procedure, image is input to connected component extractions to apply many things such as binarization (to isolate the text from the background), CC- black layer (show black components), CC- white layer (show white components), CC- black layer filter, CC- white layer filter. After that, apply strict-NRG for CCs text and relaxed-NRG applied without CCs text. Finally, apply OCR to recognize text. The post-processing is: the first group of single CCs is combined to make text line, the second one is TH-OCR combine between OCR and distance, the goal is to make recognition with small distance better.

Luk´aˇs and Neumann Jiˇr´ı Matas[11]**.**

The proposed method includes two stages, the first stage consists of the following contents, the image gets in extremal regions (ER) to detect text based on a 6 channels combination (R, G, B, H, S, I) from RGB and HSIV models and only one gradient channel ∇. The other thing in this stage is the text classifier based on five features (area, bounding box, perimeter, Euler number, and horizontal crossing). The output from this stage is character and

non-character. The second stage will separate text content from non-textual content by calculating additional features such as (Hole area ratio, Convex hull ratio, number of output bounding inflexion points). Note that to recognize text, they used optical flow character.

Xu-Cheng Yin, Member, IEEE, Wei-Yi Pei, Jun Zhang and Hong-Wei Hao [16]**.**

The proposed system first applies the Maximally Stable Extremal Regions (MSERs) algorithm for removed redundant components in an image. After that apply (morphology clustering, orientation clustering, and projection clustering) to construct text. Next stage using probability to eliminate non-text regions under condition any region with high probability is removed. The final stage is classifying text by using the Ada-boost classifier.

Shangxuan Tian , Shijian Lu , Bolan Su and Chew Lim Tan [17]**.**

Proposed modified Histogram orientation gradient (HOG) by adding Co-Occurrence. The main core of this count is Co-Occurrence for oriented gradient between each pixel pair. This method uses many features that are extracted to an image. First, find the gradient magnitude L2 norm for (vertical and horizontal) computed by Sobel filter. The orientation using in this stage is (0^o and 180^o) because this leads to finding nine bins. The second part is voting by using weighting based on gradient magnitude, such as every two neighbors is vote. To construct block features using L2 normalization then concatenation all normalized block features. The recognizer of this method is the support vector machine (SVM) classifier.

Cong Yao Xiang Bai_ Baoguang Shi Wenyu Liu[18]**.**

The proposed method takes an image and then identify character using voting. Describe character using histogram features with (5×7 cells). Classification of character based on sixty-three classes (fifty-two English characters with small and big letters and ten digits and one for special) with binary classification (Valid and Invalid). Finally using these sixty-three classes as recognizers.

Jerod J. Weinman, Member, IEEE, Zachary Butler, Dugan Knoll, and Jacqueline Field[7]**.**

The proposed method for reading text. Segmentation of an image depends on simple Gaussian model due to color parameter, mean, and variance. The next stage is binarization that contains connected components (include binary component only) and text classification (during high score is classified as text). After that normalize text during probabilistic regression. The final stage recognizes text during the Semi-Markov model includes characters' regions, linguistic properties.

Shangxuan Tian , Shijian Lu, Bolan Su and Chew Lim Tan [19]**.**

The proposed method is Multi-Level MSERs. There are two lines, Bright on the Dark side and Dark on the Bright side. Each line consists of isolating the R, G, and B channels of the input image and then find Multi-Level MSERs and then do segmentation for three channels then combined them to get on the binary result.

Myung-Chu, Sung, Bongjin Jun, Hojin Chand Daijin Kiml[20]**.**

This method is to detect text only. The proposed method includes character candidate extraction this contains four steps start with ER tree construction, next Sub-Path partitioning, and Sub-Path pruning, and finally character candidate selection. Character classification by using AdaBoost classifier includes four steps start with prepared positive and negative images, then calculate the error and update weight finally check stop conditions. Character refinement this is useful when some cases that classifier fails in classify so to eliminate

(avoid) error cases using heuristic rules and then text region grouping by using another heuristic rules to make neighbor regions to gather.

Lukas Neumann and Jift Matas[21]**.**

The proposed method includes detecting MSERs features that are classified as a character and multi-character and the MSERs background is ignored. Build local text model by considering each text line as a sequence of characters described by line and has the same properties. The segmentation stage using the Gaussian mixture model (same as [15]). Finally, text recognition by using optical character recognition.

Baoguang Shi, Xiang Bai, Senior Member, IEEE, and Cong Yao [22]**.**

The proposed network is based on a convolutional recurrent neural network (CRNN) that consists of three layers: convolutional layer, recurrent layer, and transcription layer with some modification such as take convolution and max-pooling from standard deep CNN with removing a fully connected layer. Use of CRNN to recognize text without using other parts that are used in scene text recognition. In the third part of CRNN using two things per-frame prediction (distribution) and predicted sequence(text)

Yang Liu, Zhaowen Wang, Hailin Jin, and Ian Wassell[2] **.**

The proposed method takes an input image and then extracted features that only consider as text by using encoder E after that there are two lines first get in G to construct $X^{\wedge}$ and the second, get on the decoder T to predict text $y^{-}.y^{-}$ convert to true text as y. y get in (R render with factor Z^{-} to remove noise to get X^{-}. To find losses in character the not recognize compare $X^{\wedge}$ with X^{-} to find the missing character.

Wenjia Wang and Yu Shen. The proposed method includes [23]**.**

The proposed method includes two parts Path aggregation contain (pooling layer, shortcut layer, cascade layer, and path fusion layer) and feature selective module contain squeeze and Excitation (SE) block and feature map U(to create multiple segmentation).

Jinyuan Zhao, Yanna Wang, Baihua Xiao Cunzhao Shi, Jingzhong Jiang, Chunsheng Wang [24]**.**

The proposed method Attentional scene text recognition (ASTR) contains two major things, first rectified network and recognition network. Input image gets in rectified layer connect two ways with recognition layer during adversely learning. Finally, the output rectified image that recognizes text.

Jiaxin Zhang Canjie Luo , Lianwen Jin, Tianwei Wang , Ziyan Li a , Weiying Zhou [25]**.**

The proposed method is Scale-aware hierarchical attention network that contains two major things, Encoder and hierarchical attention decoder. The encoder includes a CNN network for multi-scale properties and RNN for learning semantic properties. The Decoder performs attention mechanism twice on multi-scale properties.

Qingqing WANG1, Ye HUANG, Wenjing Jia, Xiangjian He [26]**.**

The ConvLSTM consists of two parts CNN and ConvLSTM. The whole layers are convolution layer, max-pooling layer, deformable convolutional layer, fully connected layer, attention equipped a layer. The output is text only.

ZhaoyiWan,1∗ Minghang He, Haoran Chen, Xiang Bai, Cong Yao1[6]**.**

The proposed method includes Encoding to generate text features that get in three main parts, the first one is character segmentation, and the second part is order segmentation and the final part is localization map. The second and third parts are getting in element-wise multiply, then get in order map. The first part with order map is to get in element-wise multiply to generate text.

5. Discussion of the previous literature

Method [14] is very complicated because it uses many networks and still needs post-processing. Method [15] depends on filters so it's quick to recognize the text, but using the same color property to isolate text is not a good way. Method [11] is good based on features for classifying text to reduce false positives. Method [16] using AdaBoost classifier is not useful because there is the ability to use networks instead of that. Method [17] is based on voting and this is not good because the voting method in many cases fails in detecting characters, voting is very useful when its use in comparison between many methods to detect characters. Method [7] same as method [16]. Method [18] there are many sub-methods to create this method this led to problem in run time. Method [19] is very beautiful because it simulates the R, G, B so detect MSERs for each channel is very good. Method [20] works in a gray channel only less than on (gray+red+green) channels. Method [2] main component is OCR and this is a popular method for recognizing text. Method [21] is based on CNN so it is very quick. Method [22] there is a loss in character when compared text from the final result with the start of the method. Method [23] detect text-based on feature map and using CNN is better. Method [6] same as the methods [11,16] but in different components. Method [24] fast method and it is like [16, 17, 20] but with different components. Method [25] based on convolution layer without fully this is not useful because CNN works better. Method [26] based on encoding the output of this method is recognized regions that consider as a character.

6. Performance measurements

The main performance criteria is used to evaluate the recognition methods are three main measures: Precision, Recall, and F-Score[27].

- Precision, Recall. and F-Score rates are computed supported by the number of correctly detected characters (CDC) in an image.
- False Positive (FP): The regions that are not a character, but detect as a character.
- False Negatives (FN): characters that don't detect truly.

Precision Rate (P): Defined as the ratio of properly detected characters to the sum of properly detected characters and false positives.

$$P = \frac{\text{CDC(true positive)}}{\text{CDC+FP}} * 100\% \quad (1)$$

Recall Rate (R): Defined as the ratio of the properly detected characters to the sum of properly detected characters and false negatives.

$$R = \frac{\text{CDC(true positive)}}{\text{CDC+FN}} * 100\% \quad (2)$$

F-Score (F-Measure): The harmonic mean of recall(R) and precision(P) rates.

$$F = 2 * \left(\frac{P * R}{P + R}\right) \quad (3)$$

7. Comparative analysis

A comparison between all methods described in the literature review depends on the criteria that describe it in section 4. There is no perfect way to solve all challenges, so some methods solve some challenges and fail with others. The methods and their failures are noted

and clarification of the points of failure for each proposed method is listed in Table-1 and Table-3. All the methods presented in this paper address some of the challenges that are fundamental to solve the problem of text recognition.

Table 1: The Gaps of the proposed methods

Method(s)	The gaps of the previous method(s)	Advantages of method(s)
SRPN and FCN based text detector [14]	• Unable to detect curved-shape • if the proposed region generated by SRPN only contains part of a text line/word, it is hard to predict an entire boundary for such text line/word • the proposed method may not be so effective in detection with reasonable time.	• Localize text of wide scale range and estimate text scale efficiently. • Text detector in the second stage detects texts of narrow scale range but accurately. • Non-text regions are eliminated through SRPN efficiently.
two-step iterative CRF algorithm with a BP inference and an OCR filtering stage [15]	There are cases where texts are not detected, which are mainly due to: • single characters. • strong highlights • transparency of the text • the size that is out of bounds • excessive blur, and • curve baseline.	• OCR confidence is used as an indicator for identifying the text regions while, • A traditional OCR filter module only considered the recognition results.
Extremal Regions with Incrementally Computable Descriptors and OCR [16]	Many false detections are caused by watermark text embedded in each image	Complexity with O(1) is the basic thing in recognize text.
(morphological, orientation, and projection) clustering [17]	Clustering fails in: • curve text • short similar multi-line text	Much better than the state-of-the-art performance.
Co-occurrence histogram of orientation (Co-HOG), and SVM classifier [18]	Certain recognition failure: • some characters are mistakenly annotated in both ICDAR and SVT datasets. • There exists character of size 1×135 pixels in the ICDAR 2003 not detected	Co-HOG is a more powerful tool that captures spatial distribution of neighboring orientation pairs instead of just a single gradient orientation.
Probabilistic Regression Formulation, with Expectation-Maximization (EM) Fitting [20]	Many spurious detections give only one character words	Using scene context to recognize several words together in a line of text, our system gives state-of-the-art performance on three difficult benchmark data sets.
multi-level MSER technique [21]	Work better only for SVT database only.	A multi-level MSER technology that identifies the best-quality text candidates from a set of stable regions that are extracted from different color channel images.
a novel method to extract character candidates from an ER tree [22]	There is only a small region detect correctly. To solve this problem, used similarity value (just for character level rate) **disadvantages.**	The maximally stable extremal region (MSER) method has been widely used to extract character candidates, but because of its requirement for maximum stability,

		high text detection performance is difficult to obtain. To overcome this problem, we propose a robust character candidate extraction method that performs ER tree construction, sub-path partitioning, sub-path pruning, and character candidate selection sequentially.
MSER with OCR recognition [23]	method failure in detecting: • character-like objects near the text (e.g. pictographs, arrows,.) • low contrast characters that are not picked up in the initial stage.	The method runs in real time and achieves state-of-the-art text localization and recognition results on the ICDAR 2013 Robust Reading dataset.
Convolutional Recurrent Neural Network (CRNN), since it combines DCNN and RNN [24]	The CRNN-based system is still preliminary and misses many functionalities. (fail to recognize some words)	The proposed algorithm performs well in the task of image-based music score recognition, which evidently verifies the generality of it.
supervised feature learning framework for text recognition [25]	Failure in detecting: • character close to the background • character (I) is recognized as (1) • characters interference • text curve up	Method works on challenging cases where input images contain severe geometric distortion, such as text on a curved path.
Aggregation with PSENet [26]	The methods do not deal with: • noisy image • small area characters	• The Progressive Scale Expansion Network (PSENet), which can precisely detect text instances with arbitrary shapes, could be a useful and simple baseline. • Aggregation in spatial and channel of the feature map, making it possible to fully utilize feature of different scales and channels to handle arbitrary-shaped and multi-scale text instances.
ASTR: rectification network, the other is the recognition network [27]	Fail in some irregular text images.	Method achieves the performance of state-of-the-art.
The scale-aware hierarchical attention network (SaHAN)	SaHAN has some failure cases, such as text is: • severely distorted • surrounded by background noise which increases the difficulty of recognition.	That SaHAN achieves state-of-the-art performance.
FACLSTM focused attention ConvLSTM	Failure in detecting • some regular text dataset • some curved text dataset	FACLSTM performs competitively on the regular, low-resolution and noisy text images, and outperforms the state-of-the-art approaches on the curved text images with large margins.
TextScanner	Does not deal with curved characters.	TextScanner shows its superiority in recognizing more difficult text such as Chinese transcripts and aligning with target characters.

Table 2: comparing methods based on R, P and F

Author	method	Recall	Precision	F-score
Lukas Neumann and Jirı Matas	Probability and OCR	67.5%	85.4%	75.4%
Xu-Cheng Yin, Member, IEEE, Wei-Yi Pei, Jun Zhang and Hong-Wei Hao	(Morphology, orientation, and projection) clustering	63%	81%	71%
Jerod J. Weinman, Member, IEEE, Zachary Butler, Dugan Knoll, and Jacqueline Field	Region Grouping, Binarization, normalization, and probabilistic Regression	41.10%	36.45%	33.71%

Table 3: Comparing methods based on different databases.

Author	method	Different databases					
Shangxuan Tian , Shijian Lu , Bolan Su and Chew Lim Tan.	Co-occurrence histogram of orientation (Co-HOG), and SVM classifier.	ICDAR 2003	SVT	ICDAR 2003	SVT	ICDAR 2003	SVT
		88%	90.1%	89.21%	76.7%	86.22%	81.2%
Myung-Chul Sungl, Bongjin Jun2, Hojin Ch02 and Daijin Kiml	Modified MSER with MLBP (RGB)	ICDAR2011 99.3%		ICDAR2011 87.7%			
Myung-Chul Sungl, Bongjin Jun2, Hojin Ch02 and Daijin Kiml	Modified MSER with MLBP (Gray)	ICDAR2013 72.01%		ICDAR2013 87.64%		ICDAR2013 79.06%	
Myung-Chul Sungl, Bongjin Jun2, Hojin Ch02 and Daijin Kiml	Modified MSER with MLBP (RGB)	74.23%		88.56%		80.80%	
Lukas Neumann and Jift Matas	MSER detector, Local iterative, And OCR	72.4%		81.8%		77.1%	

Wenhao He Xu-Yao Zhang, Fei Yin, Zhenbo Luo , Jean-Marc Ogier, Cheng-Lin Liu	SRPN	76.98%	84.85%	80.72%
Ryo Nakao, Brian Kenji Iwana, and Seiichi Uchida	a Path Aggregation Module and a Feature Selective Module	83.03%	83.03%	79.8%

8. Conclusion

This article shows a review of methods that describe and solve scene text recognition problem. From literature review, it can be concluded that most methods solve text recognition problem, but not in a perfect solution, so scene text recognition is still a challenging field for researcher to find a new method for solving the problem and challenges. From all previous methods, the challenges that tackled in the third paragraph (Introduction) are not all solved. There is no one talk about them or solve them perfectly, so any method that solves those challenges will give a new contribution and become a new method to solve scene text recognition problem. From Table 1 we found that a large number of proposed methods failed in curved text and irregular text. Another method fails in noise text image and short similar multi-line text. And the other methods fail in regular text and these methods are not suitable.

References

[1] B. E. E. Ofek, "YW: Detecting text in natural scenes with stroke width transform," 2010.

[2] Y. Liu, Z. Wang, H. Jin, and I. Wassell, "Synthetically supervised feature learning for scene text recognition," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 435–451.

[3] C. Yi and Y. Tian, "Scene text recognition in mobile applications by character descriptor and structure configuration," *IEEE Trans. image Process.*, vol. 23, no. 7, pp. 2972–2982, 2014.

[4] M. Liao *et al.*, "Scene text recognition from two-dimensional perspective," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, vol. 33, no. 01, pp. 8714–8721.

[5] U. Sajid, M. Chow, J. Zhang, T. Kim, and G. Wang, "Parallel scale-wise attention network for effective scene text recognition," in *2021 International Joint Conference on Neural Networks (IJCNN)*, 2021, pp. 1–8.

[6] Z. Wan, M. He, H. Chen, X. Bai, and C. Yao, "Textscanner: Reading characters in order for robust scene text recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, vol. 34, no. 07, pp. 12120–12127.

[7] J. J. Weinman, Z. Butler, D. Knoll, and J. Feild, "Toward integrated scene text reading," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 2, pp. 375–387, 2013.

[8] H. Chen, S. S. Tsai, G. Schroth, D. M. Chen, R. Grzeszczuk, and B. Girod, "Robust text detection in natural images with edge-enhanced maximally stable extremal regions," in *2011 18th IEEE International Conference on Image Processing*, 2011, pp. 2609–2612.

[9] A. Gonzalez, L. M. Bergasa, J. J. Yebes, and S. Bronte, "Text location in complex images," in *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, 2012, pp. 617–620.

[10] Y. Li and H. Lu, "Scene text detection via stroke width," in *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, 2012, pp. 681–684.

[11] L. Neumann and J. Matas, "Real-time scene text localization and recognition," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 3538–3545.

[12] K. Karthick, K. B. Ravindrakumar, R. Francis, and S. Ilankannan, "Steps involved in text recognition and recent research in OCR; a study," *Int. J. Recent Technol. Eng.*, vol. 8, no. 1, pp.

2277–3878, 2019.

[13] F. Naiemi, V. Ghods, and H. Khalesi, "Scene text detection using enhanced Extremal region and convolutional neural network," *Multimed. Tools Appl.*, vol. 79, no. 37, pp. 27137–27159, 2020.

[14] W. He, X.-Y. Zhang, F. Yin, Z. Luo, J.-M. Ogier, and C.-L. Liu, "Realtime multi-scale scene text detection with scale-based region proposal network," *Pattern Recognit.*, vol. 98, p. 107026, 2020.

[15] H. Zhang, C. Liu, C. Yang, X. Ding, and K. Wang, "An improved scene text extraction method using conditional random field and optical character recognition," in *2011 International Conference on Document Analysis and Recognition*, 2011, pp. 708–712.

[16] X.-C. Yin, W.-Y. Pei, J. Zhang, and H.-W. Hao, "Multi-orientation scene text detection with adaptive clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1930–1937, 2015.

[17] S. Tian, S. Lu, B. Su, and C. L. Tan, "Scene text recognition using co-occurrence of histogram of oriented gradients," in *2013 12th International Conference on Document Analysis and Recognition*, 2013, pp. 912–916.

[18] C. Yao, X. Bai, B. Shi, and W. Liu, "Strokelets: A learned multi-scale representation for scene text recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 4042–4049.

[19] S. Tian, S. Lu, B. Su, and C. L. Tan, "Scene text segmentation with multi-level maximally stable extremal regions," in *2014 22nd International Conference on Pattern Recognition*, 2014, pp. 2703–2708.

[20] M.-C. Sung, B. Jun, H. Cho, and D. Kim, "Scene text detection with robust character candidate extraction method," in *2015 13th International conference on document analysis and recognition (ICDAR)*, 2015, pp. 426–430.

[21] L. Neumann and J. Matas, "Efficient scene text localization and recognition with local character refinement," in *2015 13th international conference on document analysis and recognition (ICDAR)*, 2015, pp. 746–750.

[22] B. Shi, X. Bai, and C. Yao, "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 11, pp. 2298–2304, 2016.

[23] W. Wang and Y. Shen, "Scene Text Detection Via Cascade FPN and Channel Enhancement," in *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)*, 2019, pp. 155–161.

[24] J. Zhao, Y. Wang, B. Xiao, C. Shi, J. Jiang, and C. Wang, "Adversarial learning based attentional scene text recognizer," *Pattern Recognit. Lett.*, vol. 138, pp. 217–222, 2020.

[25] J. Zhang, C. Luo, L. Jin, T. Wang, Z. Li, and W. Zhou, "SaHAN: Scale-aware hierarchical attention network for scene text recognition," *Pattern Recognit. Lett.*, vol. 136, pp. 205–211, 2020.

[26] Q. Wang *et al.*, "FACLSTM: ConvLSTM with focused attention for scene text recognition," *Sci. China Inf. Sci.*, vol. 63, no. 2, pp. 1–14, 2020.

[27] C. Goutte and E. Gaussier, "A probabilistic interpretation of precision, recall and F-score, with implication for evaluation," in *European conference on information retrieval*, 2005, pp. 345–359.