Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.98549379

ANALYSIS OF VARIOUS APPROACHES FOR SCENE TEXT DETECTION AND RECOGNITION

Sridhar Gujjeti
Assistant Professor, Department Of CSE, Kakatiya Institute Of Technology And Science, Warangal, Email id: gs.cse@kitsw.ac.in

Sanda Radhika
Assistant Professor, Department Of CSE, Svs Group Of Institutions, Warangal
Email id: radhika2113sridhar@gmail.com

Abstract:
Natural scene text identification is critical for extracting textual information from natural settings. Natural scene text detection systems are emerging and yielding improved detection results as deep learning technology advances. The analysis and description of the current stage of deep learning-based text methods for natural scenes in this work may be separated into two types: proposal region and semantic segmentation, and the content of these two series of associated techniques is described. Second, we give a publicly available dataset and detection metrics for scene text detection. Finally, the study in scene text identification is summarized and predicted in the expectation of providing novel areas of study for future algorithms.
Keywords: Scene text, deep learning, text detection, LSTM, EAST Algorithm

1. INTRODUCTION:

Text that appears in an image acquired by a camera in an outdoor area is referred to as scene text. The identification and recognition of scene text from camera shot images are computer vision challenges that have become increasingly relevant since smart phones with good cameras became widely available. Text in scene photographs varies in shape, font, colour, and placement. Non-uniform illumination and focus can make it difficult to recognise scene text. Scene text detection systems are emerging and yielding improved detection results as deep learning technology advances. Text detection is now used in a variety of areas, including banking, education, criminal investigation, network public opinion, and others. The traditional method of text detection, on the other hand, is heavily reliant on the characteristics of manual design, which are time-consuming, difficult, and inaccurate (Linna Li et al 2022).Scene text can appear in a multitude of backdrops, including but not limited to signs, walls, glasses, and even suspended in the air, implying that it can have any background. Some backgrounds are noisy and unpleasant in and of themselves, such as glowing billboards, see-through glasses, and walls with patterns or text strips. It is not easy to distinguish text from its background. In recent years, an increasing number of devices for gathering photographs and videos (smartphones, smart watches, high-definition surveillance cameras, etc.) have been widely employed in a variety of businesses. Massive amounts of image data are available to the public. Some significant text information, such as licence plate numbers, product introduction text in billboards, road information and direction indication text in street signs, and so on, is frequently included in this image data [Fang et al 2021]. As a result, the computer detects and recognises

the text in the image, and the text information obtained is critical in promoting the development of human-computer interaction [Kisacanin et al 2015], geographic location positioning [Barber et al 2006], real-time translation [Haritaoglu et al 2001], robot navigation, and industrial automation. However, the texts in the photos vary in size, font shape and orientation, and even overlap and contamination, making text detection challenging. As a result, word extraction in natural situations has steadily become a research hotspot in image processing (Weiwei Sun et al 2022).

1.1 Text detection:

Text detection is the process of detecting the text present in the image, followed by surrounding it with a rectangular bounding box. Text detection can be carried out using image based techniques or frequency based techniques. In image based techniques, an image is segmented into multiple segments. Each segment is a connected component of pixels with similar characteristics. The statistical features of connected components are utilized to group them and form the text. Machine learning approaches such as support vector machine and convolutional neural networks are used to classify the components into text and non-text.In frequency based techniques, discrete Fourier transform (DFT) or discrete wavelet transform (DWT) are used to extract the high frequency coefficients. It is assumed that the text present in an image has high frequency components and selecting only the high frequency coefficients filters the text from the non-text regions in an image.

1.2 Word recognition

The text is considered to be identified and located in word recognition, and the rectangular bounding box containing the text is available. The term included within the border box must be recognised. The methods available for doing word recognition can be divided into two categories: top-down and bottom-up approaches.

Top-down techniques use a list of words from a dictionary to choose which word best fits the supplied image.[8][9][10] Most of these methods do not segment images. As a result, the top-down approach is also known as segmentation-free recognition.

Bottom-up techniques divide an image into several components, which are then fed via a recognition engine.[11][12][13] To recognise the text, either an off-the-shelf Optical character recognition (OCR) engine [14][15][16] or a custom-trained one is utilised.

Figure 1:Examples of scene text detection problems. Different orientations; different languages; different colours; different sizes; e complex background; f occlusion; g blur; h noise; i non-uniform illumination (Xiyan Liu et al 2019)

2.Scene text detection and recognition methods based on deep learning

In this section, we will discuss recent improvements in deep learning-based scene text detection and identification approaches. Most approaches in the deep learning era use deep learning-based models, and most academics address the topic from a variety of perspectives.

The generic multi-class target detection model is adapted into a single-class (text) detection model using the generic target detection network as the core model. Examples include RCNN [Girshick et al 2013], Faster-R CNN [] Ren S et al 2017], SSD[Liu W et al 2016], 3], YOLO [Redmon J et al 2016], and other deep neural network-based image feature extraction and detection methods. CTPN [Zhi T et al 2016], SegLink [6, ABCNet [Simonyan et al 2017], and others are examples of representative algorithms.

2.1 CTPN [Zhi T et al 2016]

Tian et al. presented CTPN [Zhi T et al 2016], which considers text sections to be sequences made up of several component connections and employs RNNs to extract sequence-encoded features for regression prediction. Figure 2 depicts the network model structure in detail. The CTPN [Zhi T et al 2016] employs VGG16 [Simonyan et al 2017] for feature extraction and Faster-RCNN [Ren S et al 2017] for anchor regression, allowing the RPN to detect multi-sized objects with a single-sized sliding window and innovate on parameters. However, CTPN has some drawbacks, including the inability to detect non-horizontal text[17].

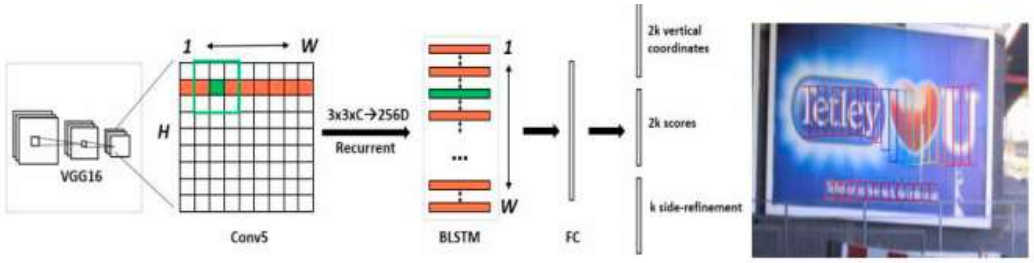

Figure 2: Network model of CTPN

2.2 PSENet [Li X et al 2016]:

Li et al. introduced PSENet, displayed in Fig. 3, a unique algorithm capable of localising text of variable shape, as well as a progressive scale expansion approach to detect neighbouring text instances. PSENet employs ResNet [He K et al 2016] as the backbone network for feature extraction and fusion, concatenates low-level texture features with high-level semantic features, predicts different kernels to obtain segmentation results for the largest kernel, and finally employs a progressive scale expansion algorithm for scaling to obtain final detection results[17].

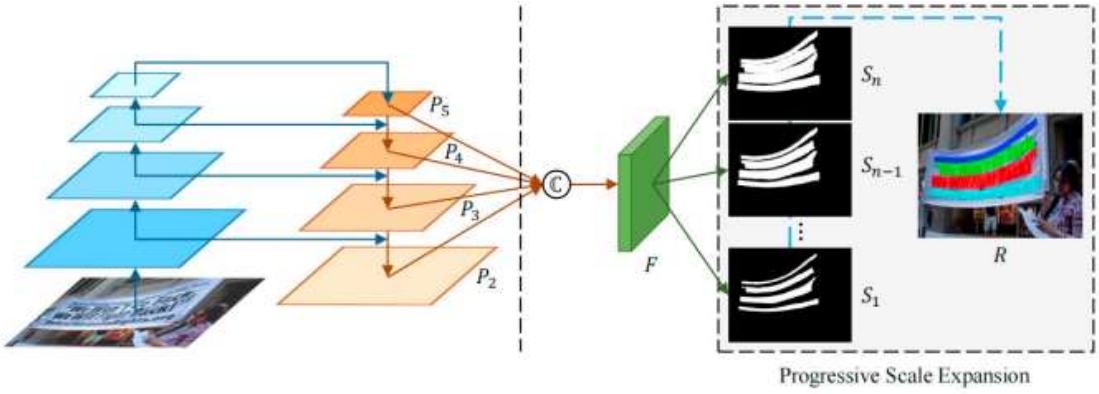

Figure 3:PSENet's network structure

2.3 SegLink [Shi B et al 2017] :

Shi et al. presented SegLink, a method based on CTPN that can detect text from any angle. In the SSD method, the former idea of employing small-scale candidate frames is combined with the latter idea of prediction based on multi-scale feature maps. SegLink's main idea is to first detect parts of the text line, then fuse the information of each picture feature if all segments are recognised, and then connect them to produce a complete text line; its network model structure is detailed in Fig. 4. [17]

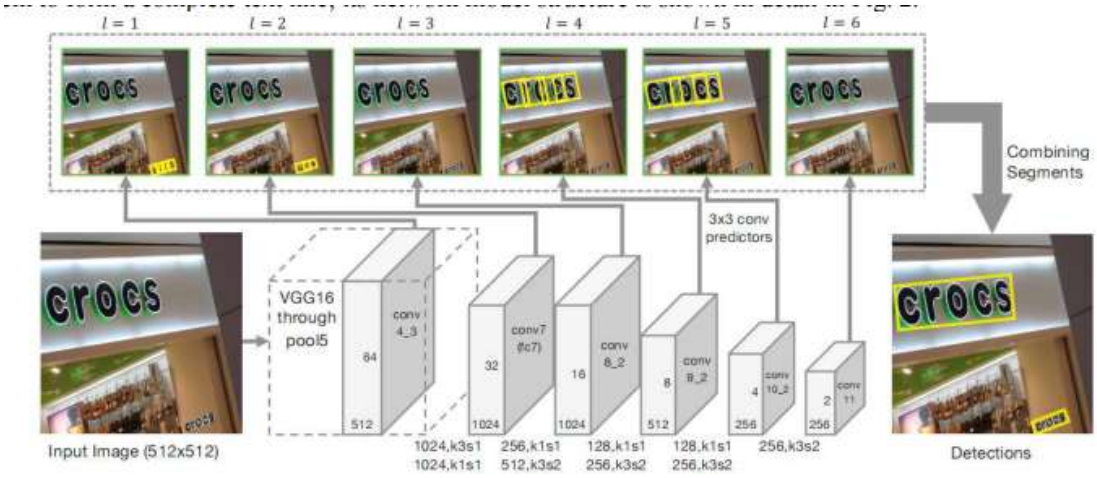

Figure 4. SegLink's network structure

2.4 ABCNet [Liu Y et al 2020]:

ABCNet [] Liu Y et al 2020], an end-to-end architecture proposed by Liu et al., is seen in Fig. 5. Bezier curves are used to detect curved text, while BezierAlign is used for feature extraction and text region correction. This allows for the recognition of oddly shaped text lines in natural settings.

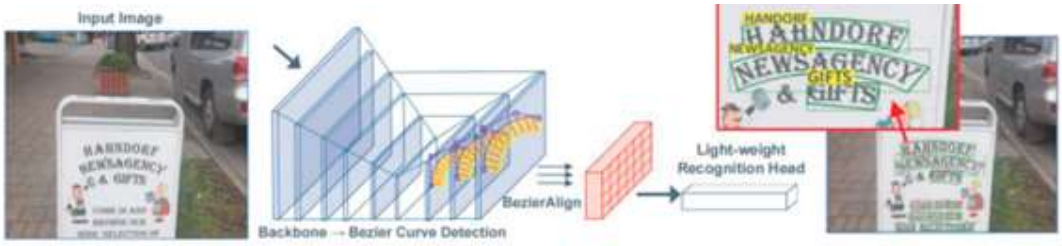

Figure 5: ABCNet network structure

2.5 EAST Algorithm: On a variety of benchmarks, scene text identification algorithms have yielded promising results. However, these methods, including those based on deep neural network models, have limitations when dealing with difficult cases. Because the overall performance of a text detection model is determined by the interactions of the algorithmic model's modules, a simple model can optimise the loss function and neural network structure in a targeted manner and improve text detection. As a result, using a simple and effective EAST technique, text regions in natural scene photos may be recognised rapidly and correctly. Figure 2 depicts the model structure. The EAST algorithm's merits include its basic structure and rapid performance. However, its text detection accuracy in complicated scenarios is inadequate**.**

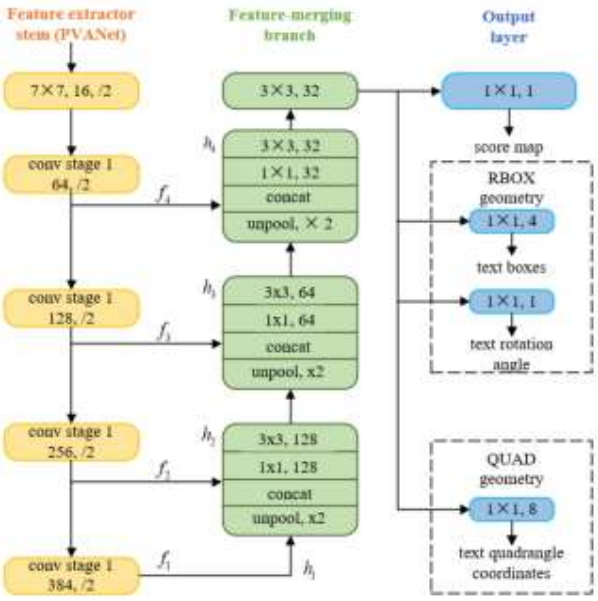

Figure 6: **EAST**

3.Contour Detection. A scene text recognition-based method is required to identify the region in the image where text is present. Instead of working on the entire image, only the object's boundary is required for further processing. Taking this consideration, the suggested approach identifies the first contour of the image [32]. Contour is used to determine the boundary of the objects present in the image. These boundaries can be identified in a variety of ways, including determining the edges of the objects and the intensities of the objects present in the image[21].

4.The Text detection is using CNN; Any model's performance is always dependent on its capacity to differentiate between different features. An image-text can be organised as a letter sequence. To recognise text in an image, a series of convolution and maxpooling layers are used. Four layers of CNN are used in the proposed method to determine whether the image's patch contains a character. ,Table 2 and Figure 4 depict the CNN configuration. First, the CNN classifier is trained with 62 classes, 26 of which are uppercase letters, 26 of which are lowercase letters, 10 of which are digits (0-9), and 1 for space. ,Because the picture patches are directly classified as letters or digits, a binary classifier is not required. The learnt characteristics are more specific and easy to distinguish from one another, making the learning process more accurate and quick. ,To detect the text in the image, a bounding box must be produced for each text present. ,This step's input image is the contour image.Because there is a potential that the input image will differ in size, the size of the input image is 24 24, and each image is a greyscale image. First, the input image is padded on all sides so that if any character is close to the image's boundary, it can be recognised using a sliding window. Each row of the image is traced using the sliding window, NMS is applied to noise if it exists in the image, and the mean deviation and standard deviation of spacing are determined. If the spacing value is less than the threshold value, it is assumed that neighbour pixels are linked. Finally, the bounding box for each character is determined using a connected component analysis approach. [21]

S. no.	Author & year	Methodology	Dataset	Performance
1	S. Yasser Arafat et al. [12], 2020	Faster RCNN + two stream deep neural network (TSDNN)	UPTI dataset	Avg. precision = 98% R. R. = 95.20%
2	Asghar Ali Chandio et al. [13], 2020	Multiscale and multilevel features	Chars74 K and ICDAR03 datasets	Precision = 90% Recall = 91% F-score = 91%
3	Yao Qin et al. [14], 2020	Faster RCNN + BLSTM	ICDAR 2015 datasets	Precision = 89.8% Recall = 84.3% F-score = 86.9%
4	Jheng-Long Wu et al. [15], 2020	BLSTM + CNN	Corpus dataset	Macro-F1 = 72% Micro-F1 = 71%
5	S. Yasser Arafat et al. [16], 2020	(AlexNet and Vgg16) + BLSTM	UPTI dataset	Accuracy = 97%
6	Sardar Jaf et al. [17], 2019	Recurrent neural network (RNN) + BLSTM	English web treebank universal dependencies dataset	Precision = 91.43% Recall = 94.52% F-score = 92.20%
7	M. A. Panhwar et al. [18], 2019	ANN	Self-dataset	Accuracy = 85%
8	Yen-Min Su et al. [19], 2019	Contour + morphological operation + ROI	ICDAR datasets	Accuracy = 93.44% Recall = 79.16% F-score = 85.71%
9	Ling-Qun Zuo Su et al. [20], 2019	CNN + BLSTM	SVT dataset, IIIT5K dataset, ICDAR 2003 and 2015 dataset	Accuracy = 95.96% Accuracy = 98% Accuracy = 98.2% Accuracy = 91%
10	Baoguang Shi et al. [21], 2017	CRNN	SVT dataset, IIIT5K dataset, ICDAR dataset	Accuracy = 97.5% Accuracy = 97.8% Accuracy = 98.7% Accuracy = 89.6%
11	Xiaohang Ren et al. [22], 2017	Text structure component detector (TSCD)	Ren's dataset, Zhou's dataset, Pan's dataset	Precision = 82% Recall = 72% F-score = 77%
12	Xiang Bai et al. [23], 2016	Bag of strokelets + HOG	SVT dataset, IIIT5K dataset, ICDAR 2003 dataset	Accuracy = 80.99% Accuracy = 85.6% Accuracy = 82.64%
13	Mingkun Yang et al. [24], 2021	CAPTCHA system	IIIT5K, SVT, IC03 IC13, IC15, SVTP CUTE	Accuracy = 92.9%, 89.6%, 92.5%, 92.2%, 76.8%, 80%, 77.1% Precision = 90.18%, 83.34%,

Figure 7: Various deep learning methods [21]

5.Scene Text Recognition Using Combined RNN and BiLSTM. This phase is used to identify the characters included in the image. In general, recognition system performance is determined on segmentation algorithms; nevertheless, good segmentation might lead to poor recognition due to noise, variable lighting circumstances, different text sizes, rotation and illumination, and so on. Deep learning-based methods are utilised, and in this paper, we integrated RNN and LSTM to improve the recognition rate to overcome these issues. ,The image's first features are extracted.The CNN classifier is used to extract sequential features from images, and training is performed for all 63 classes listed in Section 3.2.,e feature extraction is carried out using the sliding window principle, with images that have already been recognised serving as input. ,The image is first padded with 12 pixels, and the new image's size is now 24 94. A subwindow of size 24 24 is used to partition the padded image. Each portioned patch of the image is input into the trained CNN, which collects features from the image with sizes 4 4 256 and 1000, which are the output of the fourth convolution layer and the first FC layer. ,These two feature vectors are concatenated to generate a 5096-dimensional one-dimensional feature vector. To minimise the size of the feature vector, PCA and normalisation techniques are used. The new feature vector is now 256-d in size, and it contains the image's local and global features. Following the extraction of local and global features from the image, the feature labelling process begins. The suggested method employs RNN for feature labelling. RNN is a distinct neural network that can analyse sequential inputs as well as past feature information. LSTM is combined here to make the RNN more powerful. LSTM is capable of long-term memory of contextual information. The memory cell and its link to itself, as well as

three gates that control the flow of information, comprise the LSTM. Figure 8[21] depicts a visual illustration of the LSTM.

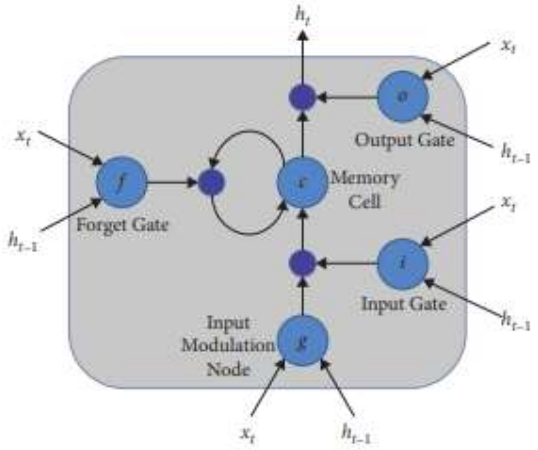

Figure 8:LSTM

6.Various datasets:

6.1 MSRATD 50 Dataset The MSRA TD dataset, which contains 3000 images of 32 32 sizes, is one of the benchmark datasets for text recognition. The dataset is difficult and noisy, including English and Chinese text. The photos in the dataset have distorted and uncertain backgrounds.Figure 6 shows the sample input photos and the recognised text, as well as its bounded box x-coordinate, y-coordinate, width, and height. The proposed method showed an accuracy of 95.22% and a recall of 85.73%. The precision is 94.15%, while the F-score is 87.09%.[21]

6.2 SVHN Dataset. The SVHN dataset (Street View House Numbers) contains 600,000 digital numbers recorded from various perspectives from various Google Street View residences. Every image are 32 32 in size, are blurred, and were taken from a different angle. The outcome's accuracy is 92.25%, and the recall, precision, and F-score are, respectively, 79.03%, 92.49%, and 89.80%[21].

A hybrid methodology employing MSER and stroke feature transform, as well as feature classification with Deep convolution neural network, is suggested for detecting text in natural photos. MSER and the stroke feature transform are used to extract the candidate character region from the image. Then, to categorise features, a Deep convolution neural network is utilised to extract deep high level features, which are merged with fully connected layers. On four benchmark datasets, SVT, ICDAR 2011, ICDAR 2013, and ICDAR 2015, the suggested technique achieves F-measures of 0.73, 0.886, 0.889, and 0.885, respectively[22].

Dataset SVT Google Street View was used to build the Street View Text (SVT) dataset. There are 101 training photos and 249 test images in total. This dataset's image is of very low quality and resolution. It lacks character boundary boxes and only includes annotations at the word level. The photos' backgrounds are quite complex. The dataset is quite difficult. Table 1 summarises the performance of various techniques on the SVT dataset. Figure 4 depicts several photos from the SVT Dataset.

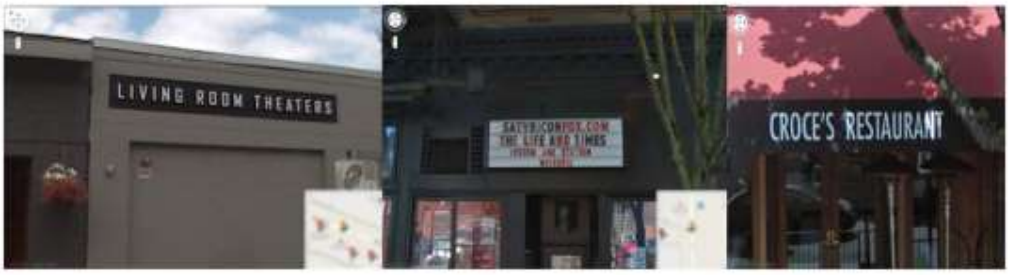

Figure 9: Sample images in SVT dataset[22].

Table 1 : lists out the performance of different approaches on ICDAR 2013 dataset[22]

Approaches	Precision	Recall	F-measure
Neumann and Matas[16]	0.191	0.329	0.242
Tang and Wu[17]	0.299	0.407	0.245
He et al[18]	0.588	0.762	0.664
Proposed method	0.598	0.78	0.73

6.3 ICDAR 2013 The most used dataset for scene text identification is the ICDAR 2013 Dataset14. There are 259 training photos and 255 testing images in all. It has 716 annotated text sections with mainly horizontal text. Table 3 illustrates the performance of various techniques on the ICDAR 2013 dataset[22].

6.4 ICDAR 2015 There are 1670 pictures and 17 548 labelled regions in the dataset. 1500 of these photos are freely accessible to the public. This is a huge dataset that is regularly utilized. The training set has 1000 photos, whereas the test set contains 500. Six institutions around the world used Google Glass devices to collect the data. The text in the photos is smaller, and the text font sizes vary. They are also present in many real-time difficulties such as noise, blur, occlusion, perspective distortion, and so on. The Dataset is a more difficult one. Table 4 lists the performance metrics for various techniques using the ICDAR 2015 dataset [22].

Table 2: Performance measures of various approaches with ICDAR 2015 dataset

Approaches	Precision	Recall	F-measure
He et al[18]	0.76	0.54	0.63
Zhang et al[14]	0.71	0.43	0.54
Wang et al[19]	0.857	0.741	0.795
Proposed method	0.859	0.758	0.885

Figure 10:Text detection in ICDAR 2013 dataset[22].

7. Conclusion:

Scene text detection is a new research area in computer vision and pattern recognition that has both theoretical and practical implications. This paper discusses current deep learning-based natural scene text detection methods, as well as their performance metrics and common datasets. The technology of natural scene text detection will continue to improve and study for even better technical solutions as computer vision and deep learning advance.

8. REFERENCES:

1. Fang, S.; Xie, H.; Wang, Y.; Mao, Z.; Zhang, Y. Read like humans: Autonomous, bidirectional and iterative language modeling for scene text recognition. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 7098–7107.

2. Wu, F.; Zhu, C.; Xu, J.; Bhatt, M.W.; Sharma, A. Research on image text recognition based on canny edge detection algorithm and k-means algorithm. Int. J. Syst. Assur. Eng. 2022, 13, 72–80. [CrossRef]

3. Kisacanin, B.; Pavlovic, V.; Huang, T.S. Real-Time Vision for Human-Computer Interaction; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2015.

4. Barber, D.B.; Redding, J.D.; McLain, T.W.; Beard, R.W.; Taylor, C.N. Vision-based target geo-location using a fixed-wing miniature air vehicle. J. Intell. Robot. Syst. 2006, 47, 361–382. [CrossRef]

5. Haritaoglu, I. Scene text extraction and translation for handheld devices. In Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Kauai, HI, USA, 8–14 December 2001; p. II.

6. DeSouza, G.N.; Kak, A.C. Vision for mobile robot navigation: A survey. IEEE Tran. Pattern Anal. 2002, 24, 237–267. [CrossRef]

7.Weiwei Sun 1 , Huiqian Wang 1 , Yi Lu 1 , Jiasai Luo 1 , Ting Liu 1 , Jinzhao Li. (2022). Deep-Learning-Based Complex Scene Text Detection Algorithm for Architectural Images. *mdpi*. 10(3914), pp.p2-22.

8. *Weinman, J.J.; Learned-Miller, E.; Hanson, A.R. (2009). "J. J. Weinmann, E. Learned-Miller, and A. R. Hanson. Scene text recognition using similarity and a lexicon with sparse belief propagation. IEEE Trans. PAMI, 31(10):1733–1746, 2009". IEEE Transactions on Pattern Analysis and Machine Intelligence. **31** (10): 1733–1746. doi:10.1109/TPAMI.2009.38. PMC 3021989. PMID 19696446.*

9. *"A. Mishra, K. Alahari, and C. V. Jawahar. Scene Text Recognition using Higher Order Language Priors. In Proc. BMVC, 2012" (PDF).*

10. *Novikova, Tatiana; Barinova, Olga; Kohli, Pushmeet; Lempitsky, Victor (2012). "Large-Lexicon Attribute-Consistent Text Recognition in Natural Images". Computer Vision – ECCV 2012. Lecture Notes in Computer Science. Vol. 7577. pp. 752–765. CiteSeerX 10.1.1.296.4807. doi:10.1007/978-3-642-33783-3_54. ISBN 978-3-642-33782-6.*

11 .*Kumar, Deepak; Ramakrishnan, A. G. (2012). "Power-law transformation for enhanced recognition of born-digital word images". D. Kumar and A. G. Ramakrishnan. Power-law transformation for enhanced recognition of born-digital word images. In Proc. 9th SPCOM, 2012. pp. 1–5. doi:10.1109/SPCOM.2012.6290009. ISBN 978-1-4673-2014-6. S2CID 13876092.*

12. *D. Kumar; M. N. Anil Prasad; A. G. Ramakrishnan. "MAPS: Midline analysis and propagation of segmentation". Proc. 8th ICVGIP, 2012. doi:10.1145/2425333.2425348. S2CID 13303734.*

13. *"D. Kumar, M. N. Anil Prasad, and A. G. Ramakrishnan. NESP: Nonlinear enhancement and selection of plane for optimal segmentation and recognition of scene word images. In Proc. 20th DRR, 2013" (PDF). 2013. doi:10.1117/12.2008519. S2CID 13848101.*

14. Abbyy Fine Reader. http://www.abbyy.com/

15 .Nuance Omnipage Reader. http://www.nuance.com/

16. Tesseract OCR Engine. http://code.google.com/p/tesseract-ocr/

17.Lingqian Yang, Daji Ergu*, Ying Cai*, Fangyao Liu, Bo Ma. (2022). A review of natural scene text detection methods. *Procedia Computer Science*. 199(2), p.p1458–1465.

[18] Cong Y, Xiang B, Liu W, Yi M, Tu Z. Detecting texts of arbitrary orientations in natural images// Computer Vision & Pattern Recognition. IEEE, 2012.

[19] Nayef N, Fei Y, Bizid I, Choi H, Ogier JM. ICDAR2017 Robust Reading Challenge on Multi-Lingual Scene Text Detection and Script Identification - RRC-MLT// 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). IEEE, 2017.

[20] Gomez R, Shi B, Gomez L, Numann L, Karatzas D. ICDAR2017 Robust Reading Challenge on COCO-Text// 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). IEEE Computer Society, 2017

21. MVV Prasad Kantipudi , 1 Sandeep Kumar , 2 and Ashish Kumar Jha. (2021). Scene Text Recognition Based on Bidirectional LSTM and Deep Neural Network. *Computational Intelligence and Neuroscienc*. -(-), pp.p1-8.

22.M. Vidhyalakshmi1 S. Sudha. (2019). Text detection in natural images with hybrid stroke feature transform and high performance deep Convnet computin. *wileyonlinelibrary.com/journal/cp*. -(-), pp.p1-8.