End-to-End Analysis for Text Detection and Recognition in Natural Scene Images

Ahlam Alnefaie[1], Deepak Gupta[2], Monowar H. Bhuyan[3], Imran Razzak[4], Prashant Gupta[5], Mukesh Prasad[1]

[1]Center for Artificial Intelligence, School of Computer Science, FEIT, University of Technology Sydney, Australia
[2]Department of Computer Science, National Institute of Technology, Arunachal Pradesh, India
[3]Department of Computing Science, Umea University, Sweden
[4]School of Information Technology, Deakin University, Geeloing, Australia
[5]Amity School of Engineering and Technology, Noida, India

***Abstract*—Right from the very beginning, the text has vital importance in human life. As compared to the vision-based applications, preference is always given to the precise and productive information embodied in the text. Considering the importance of text, recognition, and detection of text is also equally important in human life. This paper presents a deep analysis of recent development on scene text and compare their performance and bring into light the real modern applications. Future potential directions of scene text detection and recognition are also discussed.**

***Keywords*—**Text detection, text recognition, character recognition, natural image.

I. INTRODUCTION

Vision-based application is an excellent source of information for human-computer interaction [1], image search [2], and robot navigation [3]. In general, information becomes more beneficial to human beings when embodied in the text, especially to access and utilize textual information in images and videos. The textual information presented in videos and images can be studied by detecting and recognizing text. Also, reading and localizing text in natural scenes are not quite easy tasks. Some common challenges associated with the scene text detection are complexity, interference, diversity, noise, and distortion [4]. Scene text detection and recognition process involve five steps overall that are text detection, text localization, text tracking, segmentation, and text recognition. The main objective of text detection and text localization is to create a bounding box around the text appearing in image or video. The data flow diagram for the basic text detection and recognition system is presented in Fig. 1.

Complex backgrounds are difficult to detect and recognize. Sometimes a written text in the regular font on complex backgrounds becomes unreadable, as shown in Fig. 2. Some examples are natural scenes, including bricks and grass, in the background. Interference factors, including non-uniform illumination, noise, low-resolution images, blur backgrounds or text, and distortion, create challenges for scene text detection and recognition. Scenes text detection and recognition are not tricky if font styles and sizes are the same in the document. However, in the presence of different font's sizes, fonts, scales, and colors, detection and recognition process takes time and sometimes comes up with the wrong outcomes. It has been observed that sometimes images are blurred, and there is too much noise as well as a distortion in the image, which means that it is hard for the detection technology to detect and recognize correct text from the images. If the process has to be more advanced in this regard that it may also detect blur words with noise in the image, then various explorations of the process can be made, but still, it would remain to be a big challenge. The main challenge in text detection is developing a system which is able to deal with different conditions of text such as fonts, sizes, shading, lighting, and aliasing [5].

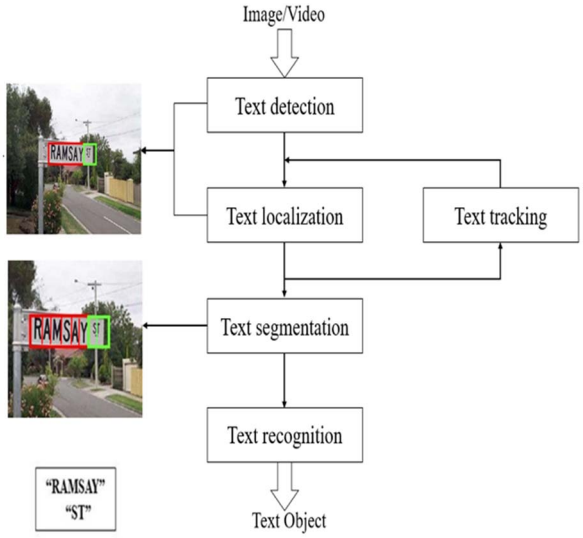

Figure 1 : Illustration of the basic scene text detection and recognition process

Figure 2 : Example of unreadable text image

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

II. Related work

Several scene text detection and recognition methods have been studied with a focus on the multi-oriented and diverse text. Most of the work is application-specific, and there is still a need to work in developing domain-independent systems. Dai et al. [6] present a scene text detection for multi-oriented text. This method contains three functions: feature extraction, feature fusion, and text prediction, which is based on Mask-NMS to obtain final results of text detection. Through flittered overlapped text by NMS mechanism, the platform conserved highest accuracy.

Liao et al. [7] propose a scene text detector that is called Textboxes. This detector is trainable and fast. Also, it detects text with high accuracy and efficiency in a single network. Textboxes method has three tasks, which are text detection, word-spotting, and end-to-end recognition. This system has three datasets that are Synth dataset for pre-training, ICDAR 2011 (IC11) for evaluating, ICDAR 2013 (IC13) for training and Street View Text (SVT) for word spotting. Textboxes method cannot detect the multi-oriented texts. Bartz et al. [8] design a step towards semi-supervised neural networks for text recognition. They proposed using a single deep neural network for text detection and recognition. This network learns to detect text from images in a semi-supervised technique. The experiment has three datasets, which are ICDAR 2013, SVT dataset, and French Street Name Signs (FSNS) dataset. The limitation of this approach is on detecting text in arbitrary locations in the images.

Ma et al. [9] develop a framework based on Rotation Region Proposal Networks (RRPN). The proposed approach is built to detect arbitrary-oriented text from natural scene images. This framework used Rotation Region-of-Interest (RRoI) to classify a text region from a feature map. Experimental datasets are MSRA-TD500, ICDAR2013, and ICDAR2015. The performance of text detection of this system is effectiveness and efficiency. Moysset et al. [10] propose a system for analyzing a full-page document. Murder dataset was used for French and English languages. A 2D-LSTM based recognizer is improved by adding the label of End-of-line (EOL). Also, this recognizer is trained with CTC alignment. Liu et al. [11] combine the two tasks of text recognition and detection as one process, through using RoIRotate operation. Their platform uses evaluation benchmarks, which are ICDAR 2015, ICDAR 2017 MLT, and ICDAR 2013 datasets. Also, their study is focused on developing a real-time oriented text recognition system on ICDAR 2015. Bai et al. [12] propose a new technique that is called edit probability. The researchers approved that an EP method can effectively control both misalignment problem and missing characters.

Deng et al. [13] introduce a detection algorithm called PixelLink. The main idea of PixelLink is based on instance segmentation, by linking pixels. The segmentation-based method can effectively detect better than regression-based methods. Li et al. [14] solve two problems of text detection, texts with arbitrary shapes and text that are very close to each other. They proposed a segmentation-based detector called Progressive Scale Expansion Network (PSENet). The datasets for evaluation that have been used are ICDAR 2015, ICDAR 2017 MLT, and SCUT-CTW1500. This method is robust for arbitrary shapes text and texts that are close to each other. This method is robust for arbitrary shapes text and texts that are close to each other. The following section has more significant studies of recent researches that have conducted this direction of study.

III. Significant improvements studies

This section produces more recent studies regarding text detection and recognition. Each of the selected studies made improvements in the field of text detection and recognition. Cheng et al. [15] propose a method known as FAN to observe the poor performance of an existing attention network method of scene text recognition on low-quality and complicated images. For comprehensive performance comparison, they compared the performance of text recognition of their method FAN with AN method. Fig. 3 illustrated the results of text recognition of FANand AN methods with green characters as correctly recognized characters**.** They observed the effect of super-parameter λ value on the accuracy results on unconstrained benchmarks. Their method produced a higher performance with $\lambda = 0.01$ and increasing the ratio of pixel-labelled samples from 0 to 30%.

Figure 3 : The output of text recognition of FAN and AN methods [15]

Tian et al. [16] present a new technique T2DAR in order to produce improved results on the performance of text detection and recognition from web videos by a unified framework based on text tracking. They used the video text dataset that is USTB-VidTEXT for assessment with sequence video frames. The result of their method is shown in Table 1. Their approach has some issues regarding text tracking when different text areas are assigned to one ID when the frames have similar locations, similar scales, and the same backgrounds.

Borisyuk et al. [17] develop an optical character recognition system to process images at Facebook scale. Their approach based on Faster-RCNN to detect words,

a fully-convolutional CNN, and CTC loss to improve the accuracy of word recognition. They used Levenshtein's edit distance to solve the issue of misinformation about the incorrect transcriptions for measuring the performance during the evaluation step. The result is improved with lower edit distance, as shown in Table 2.

Table 1 : Performance evaluation of T2DAR method [16]

Sequence name	Precision	Recall	F-score
Curtains	0.5285	0.5531	0.5405
Zippers	0.6158	0.6025	0.6091
Wood Box	0.5906	0.5892	0.5899
Biscuit Joiner	0.7000	0.7007	0.7003

Table 2 : Performance evaluation of Borisyuk et al. model [17]

Model	Training	RelativeAccuracy	Edit Distance
CHAR	Synthetic	+0.0%	-0.0%
CTC	Synthetic	+6.76%	-11.23%
CHAR	Synthetic → Human rated	+42.19%	-67.01%
CTC	Synthetic → Human rated	+48.06%	-78.17%

Liu et al. [18] report a new method DMPNet based on CNNs to address the issue of multi-orientation text and detect the text with tighter quadrangle with improving the recall rate. They solved the issue of background noise that occurs from a small threshold, which is decided using a quadrilateral sliding window. Also, recall and precision have been improved when the threshold reaches a higher ratio.This approach achieved 70.64% F. Fig. 4 presents the detected results taken from the test set of ICDAR 2015 challenge 4.

Shi et al. [19] propose an end-to-end trainable framework CRNN that handles sequences recognition in arbitrary length without character segmentation nor normalization. They adopted rectangular pooling windows instead of the conventional squared ones to improve the recognizing step for English texts and characters that have narrow shapes. The performances of this model are remarkable in lexicon-free and lexicon-based scene text recognition tasks. The recognition accuracy on IC03 dataset is 95.5%, as shown in Fig. 5. They recognized the impact of parameter δ in the accuracy, and larger δ results in more accurate lexicon-based transcription. The CRNN method achieved the maximum accuracy of 97.6% based on text set from IIIT5k dataset. Shi et al. [20] report a text recognizer with the feature of automatic rectification to enhance the recognition of random text, perspective text, and curved text. The proposed system contains three techniques that are a spatial transformer network, sequence recognition network, and TPS transformation to adapt a sequence recognition approach. They solved the problem of recognizing test with very large lexicons and the time consuming that requires iterating over all lexicon words, through adopting an efficient search scheme. However, this model has an issue when the curve angles of the text are too large. The recognition result is demonstrated in Fig. 6.

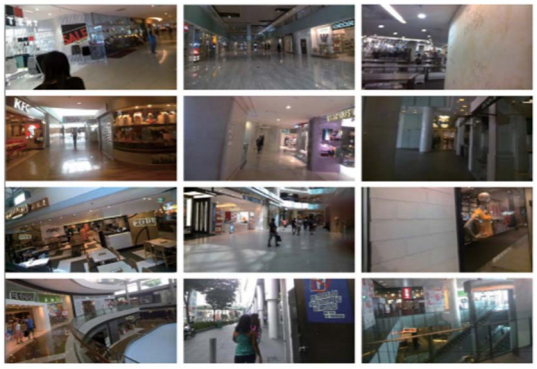

Figure 4: Experimental results of DMPNet [18]

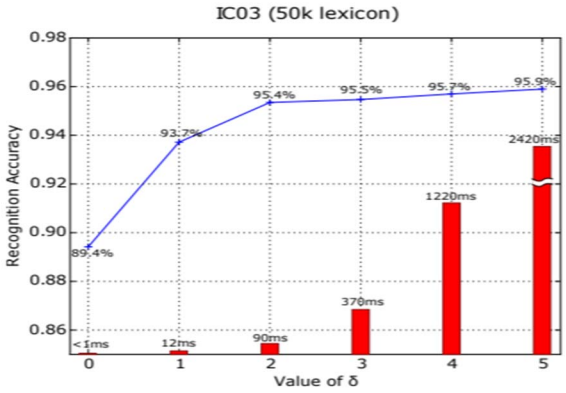

Figure 5 : Recognition accuracy of CRNN method [19]

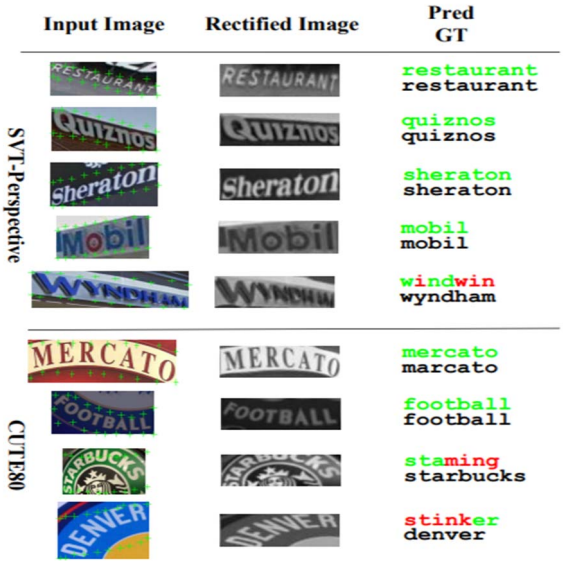

Figure 6 : Recognition accuracy of Shi et al. method [20]

Krishnan et al. [21] develop a deep convolutional neural network based method for handwritten images recognition and word spotting. The evaluation based on IAM dataset and historical document. The issue of text and image representations lie close to each other has been

solved by using the word attribute framework and embedding the label information into a commonly reduced subspace. The rate of performance on word spotting is 91.58% of mAP, and the mean word error on the word recognition is 6.69%. Fig. 7 presents recognized text with a few challenging of word images from the IAM dataset.

Figure 7 : Word recognition results from Krishnan et al. method [21]

Kumar et al. [22] present a study of recognition 3D texts that drawn by fingers and air writing through using Leap motion sensor. They increased the accuracy of words and characters detecting by using feature point extraction, cosine similarity and Ecludien algorithm. The accuracy of this 3D text recognizer is 86.88% and 81.25% using BLSTM-NN and HMM classifiers. Word segmentation accuracy that has been achieved is 78.2%. Also, some text has not been correctly recognized because the writing style of each writer was different, and the words consist of single stroke only based on 3D space recorded, as shown in Table 3. They improved the word recognition rate by combining HMM and BLSTM datasets that involve some trajectories that are correctly recognized. Table 4 presents the recognition performance of HMM and BLSTM.

Table 3 : Example of wrong transcription [22]

S.No.	3D TRAJECTORIES	ORIGINAL TEXT	RECOGNIZED TEXT
a.		dream	dress
b.		blue	navy
c.		gone	blue
d.		access	dress

Table 4 : Recognition results [22]

S.No.	3D TRAJECTORIES	ORIGINAL TEXT	BLSTM-NN RECOGNITION	HMM RECOGNITION
a.		access	X across	✓ access
b.		conquer	✓ conquer	X cover
c.		easy	✓ easy	✓ easy
d.		world	X work	X dress

Zhong et al. [23] implement a novel unified framework DeepTextfor text region proposal generation and text detection based on a fully convolutional neural network. They applied the filtering algorithm and an iterative bounding box voting scheme, to remove redundant boxes for each text instance, also to enhance the performance of the text detection step. This approach uses Inception-RPN and a set of text characteristics to improve high word recall with hundred level candidate proposals. Table 5 shows the performances of this framework evaluated on CDAR 2011 and ICDAR 2013.

Table 5: Performances evaluation ofDeepText method [23]

Dataset	Precision	Recall	F-score
CDAR 2011	0.85	0.81	0.83
ICDAR 2013	0.87	0.83	0.85

IV. BENCHMARK DATASETS AND PERFORMANCE EVALUATION

In this section, we explain the recently advanced datasets for text detection and recognition evaluating process. Comparison of the existing datasets with different factors is summarized in Table 6. The progress of recent algorithms in scene text detection and recognition has been achieved by the datasets and assessment processes in these fields. The main three metrics in performance evaluation of scene text detection methods are precision, recall and F-score. Precision is the ratio between the actual positives and all detections. The recall is the ratio between the actual texts that should be detected, while F-score is an accurate indicator of algorithmic performance.

Table 6 : Benchmark datasets

Dataset	Annotation	Orientation	Language
COCO-TEXT [24]	Word	Horizontal	English
EMINST [25]	Handwritten Character	Horizontal	English
MSRA-TD500 [26]	Text line	Multi oriented	English Chinese
CTW [27]	Text line	Horizontal	Chinese
ICDAR2017 [28]	Word	Multi oriented	Multi-lingual
TOTAL-TEXT [29]	Word	Curve	English
UBER-TEXT [30]	Text line	Horizontal	English
SVT [31]	Word	Horizontal	English
CHARS74K [32]	Character	Horizontal	English Kannada

Table 7 : Performance evaluation using Coco-Text [24]

Algorithm	Precision	Recall	F-score
Lyu et al. [33]	61.9	32.4	42.5
Yao et al. [34]	43.23	27.1	33.31
Liao et al. [35]	64	57	61

Coco-Text is a new large-scale dataset which is used detection and recognition of text in natural images. Veit et al. [24] evaluate Coco-Text dataset through three OCR algorithms A, B and C from their collaborators at Google, TextSpotter and VGG. The performances of different text detection methods evaluated on the Coco-Text dataset are shown in Table 7.

EMNIST is a dataset that consists of a set of handwritten character digits that are derived from NIST Special Database, and they are then converted to a format of 28x28 pixel image. Also, this dataset structure directly matches the MNIST dataset. EMNIST dataset contains two formats and six different splits. Baldominos et al. [36] review work contributions evaluated on the EMNIST dataset for handwritten text recognition. The performances of different text recognition algorithms evaluated on the EMNIST dataset with maximum accuracy score are presented in Table 8. . EASTR is multilingual Arabic, English scene text recognition dataset consisting of 42K text images [59, 60].

MSRA-TD500 database is released publicly to evaluate text detection algorithms for tracking text detection in natural images [26]. This database contains 500 original photos which are captured from indoor and outdoor using a pocket camera. The training set contains 300 images which are selected randomly from the original dataset, and the testing set contains the remaining 200 images. All the images are thoroughly annotated. The performances of different text detection methods evaluated on the MSRA-TD500 dataset are shown in Table 9. The two methods of Kang et al. [37] and Yin et al. [38] achieved significant performance in this database. Also, these methods have been used the clustering strategy for text line grouping to drive to this result.

Table 8 : Performance evaluation using EMNIST

Algorithm	Accuracy
Ghadekar et al. [39]	97.74%
Botalb et al. [40]	99.2%
Peng et al. [41]	99.75%
Singh et al. [42]	99.62%
Santos et al.[43]	99.775%
Cavalin et al. [43]	99.46%
Shawon et al. [44]	99.79%

CTW is an extensive Chinese text dataset in the wild. While there were OCR in document images which were well studied and many commercial tools were available, but there was a challenging problem for detection and recognition of text in the natural languages, especially for some more complicated characters like Chinese text [27]. CTW is a newly created dataset of Chinese dataset which contains about one million Chinese characters from 3850 unique ones annotated by experts in over 30000 street view images. This dataset includes 32,285 high-resolution images, 1,018,402 character instances, 3850 character categories, and six kinds of attributes. The character recognition accuracy for this dataset is top-1 accuracy of about 80.5%, character detection at the percentage of 70.9% and text line detect (AED of 22.1). This dataset is publicly available along with the source code and trained models.

ICDAR2017 dataset is a large multi-lingual text dataset that contains text scene images with nine languages. This dataset has 7200 training images, 1800 validation images and 9000 testing images in this dataset. The performances of different text recognition algorithms evaluated on the ICDAR 2017 dataset are presented in Table 10.

Table 9 : Performance evaluation using MSRA-TD500

Algorithm	Precision	Recall	F-score
Kang et al. [37]	0.71	0.62	0.66
Yin et al. [38]	0.71	0.61	0.66
Yao et al. [26]	0.63	0.63	0.60

Table 10 : Performance evaluation using ICDAR2017

Algorithm	Precision	Recall	F-score
Li et al. [14]	77.01	68.4	72.45
Zhong et al. [45]	75	66	70
Lyu et al. [33]	74.3	70.6	72.4
Liu et al. [11]	81.86	62.3	70.75

Synthetic word dataset is used for natural scene text recognition. It is a highly realistic dataset and sufficient to replace real data and gives us an infinite amount of training data [46]. TheSynth dataset has three models which are reading the words in an entirely different way of 90k-way dictionary encoding, the encoding of character sequence, and N-grams bag encoding. Jaderberg et al. [47] report the accuracy of their method performance that been evaluated by Synth dataset with 95.2%. It is fast, simple, and requires zero data acquisition costs. For the applications like scanning printer generated a document, the synthetic text dataset may be useful. Uber-Text is also a large scale dataset for OCR. This open-source dataset contains street-level images which are collected from car-mounted sensors [30]. It contains up to 110k images of street-side images with their text region polygons and the corresponding transcriptions. The dataset is split in training, validation, and testing subset.

SVT dataset is the abbreviation of street view text dataset. This dataset is taken from Google street view, and the images in this dataset are often in low resolution [48]. There are two characteristics which are noted in dealing with an outdoor street level image. One is that image text comes mostly from business signage, and the other is that business names are readily available through searching business geographically. These characteristics make this dataset more unique from different datasets. The performances of different text recognition algorithms evaluated on the SVT dataset are presented in Table11. The goal of SVT is to identify the words from

nearby businesses. SVT dataset has only word-level annotations. It should be used for cropped lexicon-driven word recognition and full image lexicon-driven word detection and recognition.

Chars74k dataset is used for character recognition in natural images. In this dataset, both the English as well as Kannada symbols are available. De Campos et al. [32] evaluate the accuracy of classification in different algorithms performance on cropped characters by the average recognition rate at the character level. The dataset consists of 64 classes (0-9, A-Z and a-z), 7705 characters which are obtained from natural images, 3410 characters that are hand-drawn characters and 62992 synthesized characters from computer fonts.

Table 11 : Performance evaluation using SVT

Algorithm	Accuracy
Bissacco et al. [49]	90.39%
Bai et al. [12]	87.5%
Jaderberg et al. [50]	86.1%
Bai et al. [51]	75.89%

V. RECENT APPLICATIONS OF SCENE TEXT DETECTION AND RECOGNITION

Recently, text detection and recognition in the wild have become research topics extensively. A lot of related applications, models, algorithms, and platforms have been proposed and designed [47, 52, 53]. Applications that use text recognition are often called OCR (Optical Character Recognition) software, and they convert the pictures into text. These types of software analyze the documents and compare fonts that are stored in the database. There are several OCR applications, some of them are mentioned below:

(i) Office Lens is software for smartphones developed by Microsoft, which allows you to digitize notes on chalkboards or whiteboards [54]. Digital copies of business cards, posters or documents can also be made from this app, and after that, you can also trim or edit them. Office Lens is available in the App Store and Google Play as well. Office Lens has a limited number of uploaded images per file.

(ii) OCR is also built into Google Drive. Google docs have a feature of converting images to text docs and saving in Google Drive. Text recognition through Google Drive docs has limited fonts, resolution, and size [55]. Then, document font should be the style of Times New Roman or Arial for the better results. Images can be individually processed or can be processed in multi-page PDF docs. Google docs also support a wide range of languages from Zulu, Finnish to Filipino and Yiddish. The document resolution should be at least 10 pixels in height.

(iii) The field of robotics adopts this technology. They allow robots to detect and understand the text as a human function. Researchers from the University of Pennsylvania in the GRASP laboratory developed a robot called Graspy that has the ability to read a text from indoor signs [56]. This robot can walk around, detect a word in the environment, and convert that text to speech systems. They achieved a precision of 45% and recall of 67% (harmonic mean of 54%) in just 0.18 s. Graspy has a limitation that is the ability to detect limited words with specific fonts. Case et al. [57] apply a PR2 robot platform to read the text in the office environment. Table 12 shows the recognition accuracy of their framework.

(iv) The new Anki Vector robot is smart enough to detect more than an object. Then, Anki robot can detect faces, text, answer questions, and play games too. This robot has the ability to text-to-speech and using this feature whenever that machine needs it [58].

(v) Sophiais also an intelligent robot and considered to be the world's smartest robot. Sophia has fantastic features, including facial expression and talking. Sophia also has text recognition features, including text to speech. It can talk in various languages as well, due to these features [59].

(vi) Aeolus robot is considered as a household robot for providing assisting inside the home [60]. Aeolus can recognize text commands as well as voice commands.

A lot of more robots are there in the world, and as the technology is getting more advanced, amazing products are being launched, and mostly the focus on latest technology and scientists is on computer vision in AI.

Table 12 : Text recognition accuracy [57]

Edit Dist.	Nameplate Data	ICDAR 2003
0	66%	59%
1	76%	66%
2	80%	72%

VI. CONCLUSION

In this paper, we present a selective review on the methods that have explicitly focused on improving text detection and recognition techniques. A survey of selected studies that enhance text detection methods has been reviewed, and the contribution of each study has been presented. Furthermore, a comparison has been made between recently available benchmark datasets. The development of the text detection technique is raised in the area of neutral images, however, more studies need to be conducted to improve the recognition performance of multiple words from online videos. Most of the existing systems are concerned with text in English and Chinese. Then, it is crucial to develop text detection and recognition applications that can handle texts of different languages on the same platform.

REFERENCES

[1] C. Leubner, C. Brockmann, and H. Muller, "Computer-vision-based human-computer interaction with a back projection wall using arm gestures," in *Proceedings 27th EUROMICRO Conference. 2001: A Net Odyssey*, 2001, pp. 308-314: IEEE.

[2] S. S. Tsai, H. Chen, D. Chen, G. Schroth, R. Grzeszczuk, and B. Girod, "Mobile visual search on printed documents

using text and low bit-rate features," in *2011 18th IEEE International Conference on Image Processing*, 2011, pp. 2601-2604: IEEE.

[3] D. Murray and J. J. Little, "Using real-time stereo vision for mobile robot navigation," autonomous robots vol. 8, no. 2, pp. 161-171, 2000.

[4] Y. Zhu, C. Yao, and X. Bai, "Scene text detection and recognition: Recent advances and future trends," Frontiers of Computer Science vol. 10, no. 1, pp. 19-36, 2016.

[5] Q. Ye and D. Doermann, "Text detection and recognition in imagery: A survey," IEEE transactions on pattern analysis machine intelligence vol. 37, no. 7, pp. 1480-1500, 2014.

[6] Y. Dai *et al.*, "Fused text segmentation networks for multi-oriented scene text detection," in *2018 24th International Conference on Pattern Recognition (ICPR)*, 2018, pp. 3604-3609: IEEE.

[7] M. Liao, B. Shi, and X. Bai, "Textboxes++: A single-shot oriented scene text detector," IEEE transactions on image processing vol. 27, no. 8, pp. 3676-3690, 2018.

[8] C. Bartz, H. Yang, and C. Meinel, "STN-OCR: A single neural network for text detection and text recognition," arXiv 2017.

[9] J. Ma *et al.*, "Arbitrary-oriented scene text detection via rotation proposals," IEEE Transactions on Multimedia vol. 20, no. 11, pp. 3111-3122, 2018.

[10] B. Moysset, C. Kermorvant, and C. Wolf, "Full-page text recognition: Learning where to start and when to stop," in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, 2017, vol. 1, pp. 871-876: IEEE.

[11] X. Liu, D. Liang, S. Yan, D. Chen, Y. Qiao, and J. Yan, "Fots: Fast oriented text spotting with a unified network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5676-5685.

[12] F. Bai, Z. Cheng, Y. Niu, S. Pu, and S. Zhou, "Edit probability for scene text recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1508-1516.

[13] D. Deng, H. Liu, X. Li, and D. Cai, "Pixellink: Detecting scene text via instance segmentation," in *Thirty-second AAAI conference on artificial intelligence*, 2018.

[14] X. Li, W. Wang, W. Hou, R.-Z. Liu, T. Lu, and J. Yang, "Shape robust text detection with progressive scale expansion network," arXiv 2018.

[15] Z. Cheng, F. Bai, Y. Xu, G. Zheng, S. Pu, and S. Zhou, "Focusing attention: Towards accurate text recognition in natural images," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5076-5084.

[16] S. Tian, X.-C. Yin, Y. Su, and H.-W. Hao, "A unified framework for tracking based text detection and recognition from web videos," IEEE transactions on pattern analysis machine intelligence vol. 40, no. 3, pp. 542-554, 2017.

[17] F. Borisyuk, A. Gordo, and V. Sivakumar, "Rosetta: Large scale system for text detection and recognition in images," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018, pp. 71-79.

[18] Y. Liu and L. Jin, "Deep matching prior network: Toward tighter multi-oriented text detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1962-1969.

[19] 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 transactions on pattern analysis machine intelligence vol. 39, no. 11, pp. 2298-2304, 2016.

[20] B. Shi, X. Wang, P. Lyu, C. Yao, and X. Bai, "Robust scene text recognition with automatic rectification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4168-4176.

[21] P. Krishnan, K. Dutta, and C. Jawahar, "Deep feature embedding for accurate recognition and retrieval of handwritten text," in *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, 2016, pp. 289-294: IEEE.

[22] P. Kumar, R. Saini, P. P. Roy, and D. P. Dogra, "Study of text segmentation and recognition using leap motion sensor," IEEE Sensors Journal vol. 17, no. 5, pp. 1293-1301, 2016.

[23] Z. Zhong, L. Jin, S. Zhang, and Z. Feng, "Deeptext: A unified framework for text proposal generation and text detection in natural images," arXiv 2016.

[24] A. Veit, T. Matera, L. Neumann, J. Matas, and S. Belongie, "Coco-text: Dataset and benchmark for text detection and recognition in natural images," arXiv 2016.

[25] G. Cohen, S. Afshar, J. Tapson, and A. Van Schaik, "EMNIST: Extending MNIST to handwritten letters," in *2017 International Joint Conference on Neural Networks (IJCNN)*, 2017, pp. 2921-2926: IEEE.

[26] C. Yao, X. Bai, W. Liu, Y. Ma, and Z. Tu, "Detecting texts of arbitrary orientations in natural images," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 1083-1090: IEEE.

[27] T.-L. Yuan, Z. Zhu, K. Xu, C.-J. Li, T.-J. Mu, and S.-M. Hu, "A large chinese text dataset in the wild," Journal of Computer Science Technology vol. 34, no. 3, pp. 509-521, 2019.

[28] N. Nayef *et al.*, "Icdar2017 robust reading challenge on multi-lingual scene text detection and script identification-rrc-mlt," in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, 2017, vol. 1, pp. 1454-1459: IEEE.

[29] C. K. Ch'ng and C. S. Chan, "Total-text: A comprehensive dataset for scene text detection and recognition," in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, 2017, vol. 1, pp. 935-942: IEEE.

[30] Y. Zhang, L. Gueguen, I. Zharkov, P. Zhang, K. Seifert, and B. Kadlec, "Uber-text: A large-scale dataset for optical character recognition from street-level imagery," in *SUNw: Scene Understanding Workshop-CVPR*, 2017, vol. 2017.

[31] J. Xiao and L. Quan, "Multiple view semantic segmentation for street view images," in *2009 IEEE 12th international conference on computer vision*, 2009, pp. 686-693: IEEE.

[32] T. E. De Campos, B. R. Babu, and M. J. V. Varma, "Character recognition in natural images," vol. 7, 2009.

[33] P. Lyu, C. Yao, W. Wu, S. Yan, and X. Bai, "Multi-oriented scene text detection via corner localization and region segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7553-7563.

[34] C. Yao, X. Bai, N. Sang, X. Zhou, S. Zhou, and Z. Cao, "Scene text detection via holistic, multi-channel prediction," arXiv 2016.

[35] M. Liao, Z. Zhu, B. Shi, G.-s. Xia, and X. Bai, "Rotation-sensitive regression for oriented scene text detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5909-5918.

[36] A. Baldominos, Y. Saez, and P. Isasi, "A survey of handwritten character recognition with mnist and emnist," Applied Sciences vol. 9, no. 15, p. 3169, 2019.

[37] L. Kang, Y. Li, and D. Doermann, "Orientation robust text line detection in natural images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 4034-4041.

[38] X.-C. Yin, X. Yin, K. Huang, and H.-W. Hao, "Robust text detection in natural scene images," IEEE transactions on pattern analysis machine intelligence vol. 36, no. 5, pp. 970-983, 2013.

[39] P. Ghadekar, S. Ingole, and D. Sonone, "Handwritten Digit and Letter Recognition Using Hybrid DWT-DCT with KNN and SVM Classifier," in *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*, 2018, pp. 1-6: IEEE.

[40] A. Botalb, M. Moinuddin, U. Al-Saggaf, and S. S. Ali, "Contrasting Convolutional Neural Network (CNN) with Multi-Layer Perceptron (MLP) for Big Data Analysis," in *2018 International Conference on Intelligent and Advanced System (ICIAS)*, 2018, pp. 1-5: IEEE.

[41] Y. Peng and H. Yin, "Markov random field based convolutional neural networks for image classification," in *International Conference on Intelligent Data Engineering and Automated Learning*, 2017, pp. 387-396: Springer.

[42] S. Singh, A. Paul, and M. Arun, "Parallelization of digit recognition system using deep convolutional neural network on CUDA," in *2017 Third International Conference on Sensing, Signal Processing and Security (ICSSS)*, 2017, pp. 379-383: IEEE.

[43] P. Cavalin and L. Oliveira, "Confusion Matrix-Based Building of Hierarchical Classification," in *Iberoamerican Congress on Pattern Recognition*, 2018, pp. 271-278: Springer.

[44] A. Shawon, M. J.-U. Rahman, F. Mahmud, and M. A. Zaman, "Bangla Handwritten Digit Recognition Using Deep CNN for Large and Unbiased Dataset," in *2018 International Conference on Bangla Speech and Language Processing (ICBSLP)*, 2018, pp. 1-6: IEEE.

[45] Z. Zhong, L. Sun, and Q. Huo, "An anchor-free region proposal network for Faster R-CNN-based text detection approaches," International Journal on Document Analysis Recognition vol. 22, no. 3, pp. 315-327, 2019.

[46] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman, "Synthetic data and artificial neural networks for natural scene text recognition," arXiv 2014.

[47] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman, "Reading text in the wild with convolutional neural networks," International Journal of Computer Vision vol. 116, no. 1, pp. 1-20, 2016.

[48] D. Ghai and N. Jain, "Comparative Analysis of Multi-scale Wavelet Decomposition and k-Means Clustering Based Text Extraction," Wireless Personal Communications vol. 109, no. 1, pp. 455-490, 2019.

[49] A. Bissacco, M. Cummins, Y. Netzer, and H. Neven, "Photoocr: Reading text in uncontrolled conditions," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 785-792.

[50] M. Jaderberg, A. Vedaldi, and A. Zisserman, "Deep features for text spotting," in *European conference on computer vision*, 2014, pp. 512-528: Springer.

[51] X. Bai, C. Yao, and W. Liu, "Strokelets: A learned multi-scale mid-level representation for scene text recognition," IEEE Transactions on Image Processing vol. 25, no. 6, pp. 2789-2802, 2016.

[52] K. Sheshadri and S. K. Divvala, "Exemplar Driven Character Recognition in the Wild," in *BMVC*, 2012, pp. 1-10.

[53] Y. Li, W. Jia, C. Shen, and A. van den Hengel, "Characterness: An indicator of text in the wild," IEEE transactions on image processing vol. 23, no. 4, pp. 1666-1677, 2014.

[54] *Office Lens*. Available: https://www.microsoft.com/en-au/p/office-lens/9wzdncrfj3t8?rtc=1&system-requirements=&activetab=pivot:reviewstab

[55] *OCR Google Drive*. Available: https://support.google.com/drive/answer/176692?co=GENIE.Platform%3DDesktop&hl=en

[56] *University of Pennsylvania's PR2 robot learns to read*. Available: https://phys.org/news/2011-05-university-pennsylvania-pr2-robot-video.html

[57] C. Case, B. Suresh, A. Coates, and A. Y. Ng, "Autonomous sign reading for semantic mapping," in *2011 IEEE international Conference on Robotics and Automation*, 2011, pp. 3297-3303: IEEE.

[58] D. Bohn. *The new anki vector robot is smart enough to just hang out*. Available: https://www.theverge.com/2018/8/8/17661902/anki-vector-home-robot-voice-assistant-ai

[59] J. Retto, "SOPHIA, FIRST CITIZEN ROBOT OF THE WORLD," 11/27 2017.

[60] *Aeolus robot*. Available: https://aeolusbot.com/