A NEW DEFECT DETECTION METHOD FOR IMPROVING TEXT DETECTION AND RECOGNITION PERFORMANCES IN NATURAL SCENE IMAGES

[1]Hamam Mokayed, [2]Palaiahnakote Shivakumara, [3]Marcus Liwicki, and [4]Umapada Pal
[1]Email: homammo@gmail.com
[2]Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur, Malaysia.
Email: shiva@um.edu.my
[3]Department of Computer Science, Electrical and Space Engineering, Lulea University of Technology, Email: marcus.liwicki@ltu.se
[4]Computer Vision and Pattern Recognition Unit, Indian Statistical Institute, Kolkata, India. Email: umapada@isical.ac.in

ABSTRACT

This paper presents a new idea for improving text detection and recognition performances by detecting defects in the text detection results. Despite the rapid development of powerful deep learning based models for scene text detection and recognition in the wild, in complex situations (logos or decorated components connected with text), existing methods do not yield satisfactory results. In this paper, we propose to use post-processing method to improve the text detection and recognition performance. The proposed method extracts features, namely phase congruency, entropy and compactness for the text detection results. To strengthen discriminative power for feature extraction, we explore the combination of SVM classifier and Gaussian distribution of text components to determine proper weight, which represents true text component. The weights are multiplied with the features to detect defect components though clustering. The bounding boxes are redrawn, which results proper bounding box without defects components. Experimental results show that the proposed defect detection reports satisfactory results. To validate the effectiveness of defect detection, we conduct experiments on benchmark datasets of MSRA-TD-500 and SVT for detection and recognition before and after defect detection. The result shows that the performance of text detection and recognition improves significantly after defect detection.

Index Terms— Natural scene detection, Natural scene text recognition, Gaussian distribution, Text box corrections.

1. INTRODUCTION

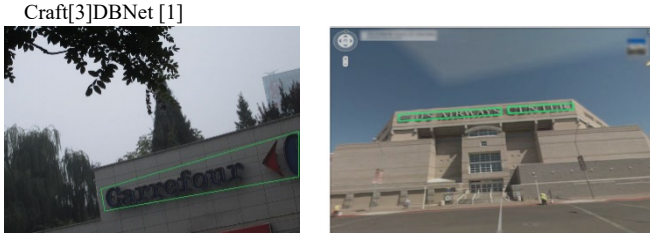
There are several situations, where the components behave like text components, such as special symbols associated with text information, when the company logos are connected with name of the buildings and when decorated component is connected to text. In these cases, the existing text detection and recognition methods (including deep learning models) fail to exclude such symbols as non-text components [1, 2]. As a result, most of the existing text detection methods include such non-text components in the text bounding boxes, leading to erroneous results for such text box and, finally, to chances of obtaining incorrect recognition results. Another reason for erroneous results is if the extracted features conflict with those non-text components due to common properties they share. This is due to the similar shape of text and non-text components. Since the text detection results contain non-text components, they are considered as defect components. This is because a defect component is a small portion, which may or may not noticeable like defects in the normal images compared to text components. It is necessary to study more local information of the non-text components to separate them from the text components. If the method identifies it as a defect component successfully, we can redraw the bounding boxes for only text components without defective portion. This helps text detection method to improve its performance significantly. If the text detection is correct, it is obvious that recognition performance improves.

It is illustrated in Fig.1, where we can notice the text detection methods, namely, CRAFT (Character Region Awareness for Text Detection) which uses character information and affinity between the characters for text detection [3], DBNet (Differentiable Binarization Network), which includes binarization step in the segmentation network for text detection [1], fail to fix proper bounding boxes for the four different situations as shown in Fig.1(a). It is evident that the symbols and decorative non-text components cause the problem for fixing improper bounding boxes, which is called defect in text detection results in this work. When we fed the defected text detection results to recognition method, namely, CRNN, which integrates feature extraction, sequence modeling and transcription into unified network [2], it outputs incorrect recognition results as shown in Fig.1(b), where character marked by red color indicates wrong recognition. With this illustration, one can infer that the existing text detection and recognition methods do not give satisfactory results for the situations shown in Fig.1(a). It motivated us to develop a new method for defect detection in the results obtained by text detection methods, which is classification of text and non-text components as shown

sample results in Fig.1(c), where one can see for all the four cases, the defects are removed successfully. These results can be used for redrawing the bounding boxes by excluding defect component as shown in Fig.1 (d). This results in enhancing the performances of text detection and recognition.

Hence, this work focus on developing a simple and effective method for detecting defect, which is non-text component, associated with detection results. This is the main contribution and is different from the existing methods. The organization of the paper is as follows. The review of existing text detection and recognition methods are presented in Section 2. The proposed method for classification of defect and text components is described in Section 3. Experimental results to validate the proposed classification are discussed in Section 4. Section 5 presents conclusions and gives and outlook to future work.

Craft[3]DBNet [1]



DBNet[1]DBNet [1]

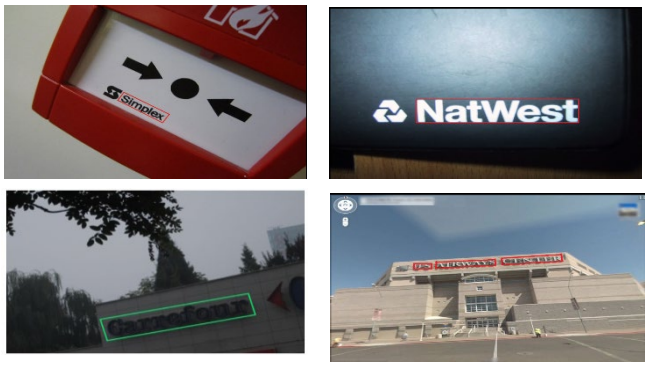
(a) Text detection results of the different methods

"5Simplex","aNatWest", "Carrefour C", "EUSAIRWAYS CENTER"

(b) Recognition of the existing CRNN method[2] for respective text detection results

"Simplex", "NatWest", "Carrefour ", "USAIRWAYS CENTER"

(c) Removing defect components from respective text detection results with the proposed method.



(d) Text detection results after removing defects for the respective four images.

Fig.1. Illustrating the need for defect detection to improve text recognition and text detection performances.

2. RELATED WORK

We review the methods on text detection and recognition in natural scene images. Most of the methods use deep learning models for achieving the results. Baek et al. [3] proposed method for text detection in natural scene images based on character region awareness. Li et al. [4] proposed method for text detection in natural scene images based on progressive scale expansion network. Liao et al. [1] proposed method for text detection in the natural scene images based on differentiable binarization network. Wang et al. [5] proposed method for scene text in the wild based on two stage network architectures. Zhu et al. [6] proposed method for scene text detection based on instance segmentation. Dai et al. [7] proposed method multi-scale context aware features aggregation for text detection in natural scene images. Liu et al. [8] proposed arbitrarily shaped scene text detection using mask tightness text detector.

There are methods that use the combination of feature extraction and deep learning for improving text detection performance. For example, Roy et al. [9] proposed a method based Delaunay triangulation for detecting text from multi-view natural scenes. Similarly, Nag et al. [10] proposed the combination of features and deep learning for detecting text in sports images. The methods are capable of detecting isolated characters but it works only for the image containing human. In the same way, Xue et al. [11] proposed arbitrarily-oriented text detection in low light natural scene images.

In summary, the methods discussed in the above are using different deep learning models for text detection in natural scene images. However, it is noted that the methods may not perform well for the text, which contains defect components, such as logos, and symbols. In addition, for isolated characters, symbols and non-text components, which look like characters associated with text, the methods do not give satisfactory results. The reason is that most of the deep learning models require context information (high level features), which represent global information of text for achieving better results. However, for single characters, isolated non-text components and if the text contain small portion of defect components, the methods lose discriminative power and hence the methods report poor results for the above mentioned situations.

For text recognition, we can find several powerful methods that use deep learning model. Shi et al. [2] proposed end-to-end deep network for scene text recognition. Carbune et al. [12] proposed LSTM deep network for online handwriting recognition. Wang et al. [13] proposed multi-branch guided attention network for scene text recognition. Zhang et al. [14] proposed scale-aware hierarchical attention network for scene text recognition. Lee et al. [15] proposed method for scene text recognition based on 2D self-attention network. Long et al. [16] proposed a method for scene text recognition using character anchor pooling. Shang et al. [17] proposed

method for scene text recognition based on character awareness network. In summary, although the methods address challenges of scene text recognition, the methods are not robust when text detection results contain non-text like symbols. Hence it is limitation of the existing text detection and recognition methods.

Thus, this work proposes a new simple and effective method to overcome the limitation of the existing methods. Motivated by angle, pixels values and shape of characters, which are prominent information for representing characters, we extract phase congruency [18, 19], average entropy [20] and compactness-based features. Inspired by the pixel distribution of characters, which generally generates Gaussian distribution, the proposed work considers the mean value obtained by the Gaussian distribution of character candidate as weight. The character candidate is detected by obtaining confidence score using SVM [21] of all the character components in the text detection results. In this work, we use simple SVM rather than CNN because the distribution clearly exhibits the difference between text and non-text components. Therefore, it does not require CNN as it may cause overfitting problem and more parameters would have to be adjusted. The weights and features are combined in a new way through clustering to separate text and non-text components, which is called defect component detection. The way the proposed method uses conventional approach to solve complex issue of defect detection to improve text detection and recognition performance is the key contribution of the proposed work.

3. THE PROPOSED METHOD

In this work, the output of text detection method is the input for defect component detection. To overcome the limitation mentioned in the introduction section, one should extract multiple local features to differentiate defect component from text components. At the same time, it is necessary to find common features among text component to group them as one cluster. It is observed that the pixel values, angles and shape of character component are the prominent features for representing text components. Motivated by these observations, we explore phase congruency, which exploits angle information of the pixels, entropy features, which exploit pixel information and compactness features, which exploit shape of the components. Due to clutter background and arbitrary orientation, defect component detection is hard. Therefore, the above-mentioned features alone may not be sufficient.

It is true that text detection results have more-number character components compared to the number of defect components. To take this advantage, the proposed approach uses SVM classifier for finding the best text candidate among text components. For the best text candidate, the proposed method finds Gaussian distribution to determine the mean value automatically, which is considered as weight. We believe that the weight value represents true text components. To increase the discriminative power of the extracted features, the weight is multiplied with those features, which widen the gap between the features of defect and text components. This gap leads to employ K-means clustering with K=2 for separating defect component from text component. The distance between the two components is estimated and the distances are supplied to K-means clustering. The cluster that contains more number of components is considered as the text components cluster. Otherwise, it is considered as defect components cluster.

3.1. Feature Extraction

For the input image, the proposed method obtains text detection results as shown in Fig.2(a), where we can see the bounding box fixed by the text detection method [3] includes non-text components, which we labeled as defect component in this work. The proposed method uses binarization approach [22] for segmenting character components from the detected text as shown in Fig.2(b). For each segmented component, the proposed method extracts phase congruency, entropy and compactness features. The phase congruency features are extracted for each pixel of segmented components. Since phase congruency involves phase angle and amplitude, it extracts angle information for each pixel. One can expect high values for the pixels which represents edge pixels and low values for background pixels. More information about phase congruency model can be found in [18, 19]. Similarly, for each segmented component, the proposed method extracts entropy features using pixel values for estimating entropy. The entropy also gives high values for the pixels, which represents edge and low values for the non-edge pixels. As mentioned earlier, shape is also important feature to separate defect components from text component, the proposed method calculates compactness for each segmented component. In this way, the proposed method extracts multiple local features for differentiating defect components from text components.

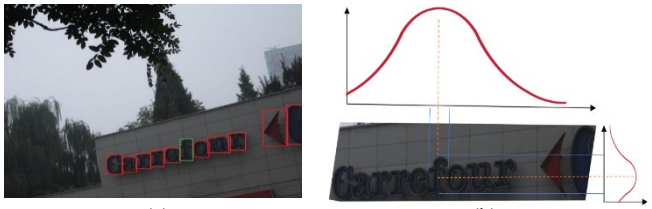
(a) (b)

Fig.2. Feature extraction for each character component region. (a) Text detection results with defect components. (b) Segmenting character components.

3.2. Defect Component Detection

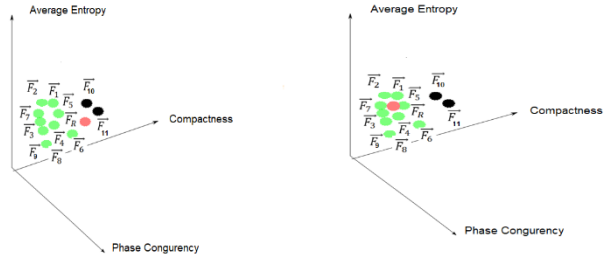
Since the considered defect component detection is challenging, the features extracted in the previous section are not sufficient. To strengthen the features, the proposed

method takes advantage of text detection results to increase the gap between features of defect and text components. It is true that text detection results usually provide more number of text components compared to the number of defect components. It can be seen in Fig.3(a), where 9 components are text components and two are defect components. This observation motivates us to detect the best text candidate among segmented components. For this, the proposed method fed all the segmented components to SVM classifier [21] to obtain confidence score for each segmented components. The component that gets highest confidence score is considered as the best text candidate as shown in Fig.3(a), where the component marked in green rectangle indicates the best text candidate.



(a) . (b)

Fig.3. Determination of weight using distribution. (a) Choosing the best text candidate among many character components. (b) Computing Gaussian distribution for chosen best candidate.



(a) (b)

(c)

Fig.4. Defect component detection in the text detection results. (a) The features distribution in 3D before multiplying weight. Green color indicates text components, black color indicates defect components and red color indicates center of the components. (b) The features distribution in 3D after multiplying weight. (c) Separating defect components from text components using clustering.

For the best text candidate, the proposed approach obtains horizontal and vertical Gaussian distribution as shown in Fig.3(b). It is noted that for any character components, the pixel distribution usually represents Gaussian shape. The same observation may not be true for defect components. With these Gaussian distributions, the proposed approach gets mean value automatically and it is considered as weight. One can understand that this weight helps us to widen the gap between the text and defect components. Therefore, the weight is multiplied with the extracted features. The effect of weight can be seen in Fig.4(a) and Fig.4(b), where we plot 3D graphs for the phase congruency, entropy and compactness features before and after multiplying weight with the features. It is observed from Fig.4(a) that the center which is marked in red color is neither close to text point, which are marked in green color nor close to defect point which are marked in black color. This shows that the features without weight do not have much difference between text and defect components. When we see the same 3D plot after multiplying weight with the feature shown in Fig.4(b), the center is moved near to text components. This results in two groups. To separate these two groups, the proposed method calculates Euclidean distance between the component and deploy K-means clustering with K=2. This results in two clusters. The cluster that has more number of components is considered as text component cluster. Otherwise, the cluster is considered as defect component cluster. In this way, the proposed method detects defect components. The new bounding box is drawn for the text components without defect components as shown in Fig.4(c). Thus, the proposed defect detection helps in improving text detection and recognition performance. The recognition result for the text in Fig.4(c) is "Carrefour" by the method [2]. The same recognition method before defect component detection reported recognition results as "Carrefour C". This is the advantage of the proposed work.

4. EXPERIMENTAL RESULTS

For evaluating the proposed defect detection, there is no standard dataset available in literature. Therefore, we choose the number of images from the benchmark datasets of natural scene images, namely, MSRA-TD-500 [9] and SVT [11] datasets. The reason to consider the above two datasets is that the MSRA was created for arbitrary text detection method evaluation, where one can expect more number of defects in text detection results. This is because fixing bounding box for arbitrarily-oriented text is not easy as fixing bounding box for horizontal text. As a result, there are high chances of fixing improper bounding box, which may include components of background and other components. Similarly, the SVT dataset consists of the images of street view, which includes name of the buildings, street names, trees etc. Therefore, text detection in such images is not easy. In addition, since the images are captured from slightly oblique angle, it affects quality of the images apart from complex background. Therefore, there are high chances of fixing improper bounding boxes for the text lines. Sample images of MSRA

and SVT datasets are shown in Fig.5, the special symbols and logos are associated with the text lines. 44 images out of 500 from MSRA and 39 images out of 350 from SVT are considered for experimentation in this work. Although, the size of the dataset is small but the images are complex to evaluate the proposed defect detection method.

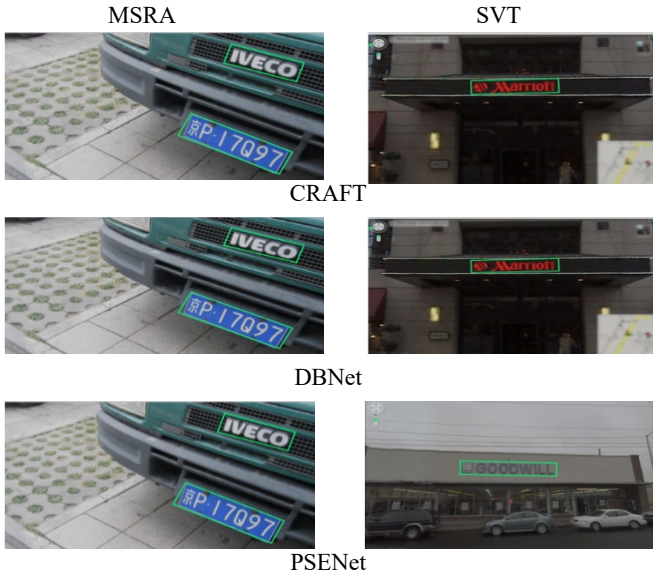


Fig.5. Text detection results of the different methods on MSRA and SVT datasets before defect detection.

To show that the proposed defect detection is effective, we implemented the following text detection and recognition methods. Baek et al. [3] method proposes Character Region Awareness for Text Detection (CRAFT), Liao et al. [1] proposes a method of Differential Binarization Network (DBNet) for text detection and Li et al. [4] proposes Progressive Scale Expansion Network (PSENet) for text detection in natural scene images. For recognition, Shi et al. [2] proposes end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition, which uses Convolutional Recurrent Neural Network (CRNN) for text recognition in the natural scene images. Carbune et al. [12] propose LSTM based method for handwriting recognition. Since the LSTM is popular for handling complex situations and it has ability to adapt for different dataset and applications, we use LSTM [12] for recognizing scene text in the images in this work. The reason for choosing the above methods to show effectiveness of the proposed defect detection is that the methods are state-of-the-art methods and the methods are capable of tackling complex situations. In addition, to show that although, the deep learning based methods are powerful in solving complex problems, they fail to obtain perfect results for some typical situations.

For measuring the performance of the proposed defect detection, we consider the standard measures, namely, Recall (R), Precision (P) and F-measure (F). The same measures are used for text detection experiments. In order to show that the proposed defect detection is effective, we conduct text detection experiments before and after defect detection by the different text detection methods. For before experiments, the input images with defect component are fed to different text detection methods for calculating measures. While for experiments after defect detection, the text detection result without defect components are supplied to the same text detection methods. It is expected that the text detection performance should be improved after defect detection significantly. For recognition experiments, we use recognition rate at word level. The recognition rate of different recognition methods before and after defect detection is calculated for the output of each text detection method. We follow the instructions mentioned in [9, 11] for calculating measures for both text detection and recognition in this work.

4.1. Evaluating Defect Detection

Quantitative results of the proposed defect detection for MSRA and SVT datasets are reported in Table 1. For the output of each text detection methods, we calculate measures for evaluating defect detection step. The results reported in Table 1 shows that the proposed method is impressive and promising as it reports reasonable good results for both MSRA and SVT datasets. When we compare the results of MSRA and SVT datasets, the results of SVT is lower than the results of MSRA in terms of F-measure for all the text detection methods results. This indicates the SVT dataset is challenging for text detection compared to MSRA dataset. This is justifiable because SVT data suffer from oblique angle while MSRA dataset does not.

Table 1. Performance of the proposed defect detection for the text detection results on different datasets

Methods	MSRA-TD-500			SVT		
	R	P	F	R	P	F
CRAFT [3]	81.4	89.2	85.12	80.6	85.8	83.12
DBNet [1]	86.3	90.1	88.16	87.6	91.8	89.65
PSENet [4]	82.4	88.9	85.53	85.7	88.4	87.03

4.2. Validating the Effectiveness of Defect Detection

Qualitative results of different text detection and recognition methods before and after defect detection are shown in Fig.6 and Fig.7, respectively. It is observed from Fig.6 that text detection method fixes proper bounding boxes for text without defect components in contrast to results shown in Fig.5. In the same way, Fig.7 shows that the recognition methods outputs correct recognition for the text after defect detection compared to before defect detection. The results on text detection and recognition show that the defect detection is essential to improve the performance of both text detection and recognition methods. Hence, the proposed defect detection is effective and useful.

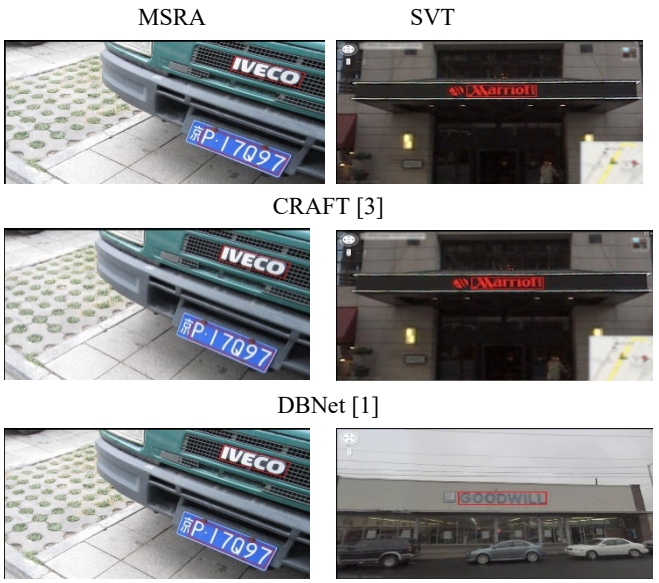


Fig.6. Text detection results of the different methods on MSRA and SVT datasets after defect detection.

Quantitative results of the text detection and recognition methods before and after defect detection for MSRA and SVT datasets are reported in Table 2 and Table 3, respectively. Table 2 and Table 3 show that the results of different text detection and recognition methods after defect detection improves significantly compared to before defect detection for both the datasets. Therefore, one can conclude that the proposed defect detection is effective and useful for enhancing the text detection and recognition performance especially for complex datasets.

“RP17Q97” “ P17Q97”

(a) Recognition result of CRNN [2] before and after defect detection

“a*P17Q97” “P17Q97”

(b) Recognition result of LSTM [12] before and after defect detection.

Fig.7. The effect of defect detection on recognition performance

Table 2. Performance of different text detection methods before and after defect detection on different datasets

Methods	MSRA-TD-500						SVT					
	Before			After			Before			After		
	R	P	F	R	P	F	R	P	F	R	P	F
CRAFT [3]	78.2	88.2	82.9	80.1	88.6	84.1	87.2	73.1	79.5	88.4	74.8	81.0
DBNet [1]	52.0	85.9	64.5	56.8	87.0	68.7	54.0	69.8	60.8	60.6	72.5	66.0
PSENet [4]	79.2	91.5	84.9	80.5	92.0	85.8	62.2	72.5	67.0	67.9	74.4	71.0

Table 3. Performance of the recognition methods before classification for the output of different text detection methods

Methods	MSRA-TD-500				SVT			
	Before		After		Before		After	
	CRNN	LSTM	CRNN	LSTM	CRNN	LSTM	CRNN	LSTM
CRAFT [3]	73.2	66.9	74.6	68.6	82.7	76.2	83.7	77.6
DBNet [1]	46.5	43.1	49.8	46.3	47.2	44.5	52.3	48.6
PSENet [4]	74.6	69.5	75.8	70.3	55.6	52.4	59.1	55.2

5. CONCLUSION AND FUTURE WORK

We have proposed a new method for defect detection in text detection results to improve the performance of text detection and recognition methods. When the text detection results include non-text components either in the beginning of text or at end of the text, such as special symbols and logos, such non-text components are considered as defect components. The proposed method extracts local information using multiple features, namely, phase congruency, entropy and compactness of each segmented character components to study angle, pixels and shape of the text components. To strengthen the above extracted features to detect defect component, the proposed approach finds the best text candidate from the segmented results with the help of SVM classifier. For the best text candidate, our method computes weight based Gaussian distribution using pixel values. The weight is multiplied with the features. This results in two clusters, namely text components cluster and defect components cluster. The results on defect dataset show that the proposed method reports satisfactory results. Furthermore, text detection and recognition experiments on two benchmark datasets show that the proposed defect detection improves the performance of the text detection and recognition methods after defect detection. However, when text detection results include defect components in the middle of text, the proposed method does not work properly. In addition, the scope of the proposed work is limited to English text and does not work on multi-lingual text. Therefore, there is a room for improvement of the proposed method in the future.

6. REFERENCES

[1] M. Liao, Z. Wan, C. Yao, K. Chen and X. Bai, “Real-time scene text detection with differentiable binarization”, In Proc. Proc. AAAI, pp 1-8, 2020.

[2] B. Shi, X. Bai and C. Yao, “An end-to-end trainable neural network for image-based sequence recognition and its applications to scene text recognition”, IEEE Trans. PAMI, pp 2298-2304, 2017.

[3] Y. Baek, B. Lee, D. Han, S. Yun and H. Lee, “Character region awareness for text detection”, In Proc. CVPR, pp. 9365-9374, 2019.

[4] W. Wang, E. Xie, X. Li, W. Hou, T. Lu, G. Yu, S. Shao, “Shape Robust Text Detection With Progressive Scale Expansion Network,”, in Proc. CVPR, pp 9328-93337, 2019.

[5] S. Wang, Y. Liu, Z. He, Y. Wang and Z. Tang, “A quadrilateral scene text detector with two-stage network architecture”, Pattern Recognition, 102, 2020.

[6] Y. Zhu and J. Du, “TextMountain: Accurate scene text detection via instance segmentation”, Patten Recognition, 2020.

[7] P. Dai, H. Zhang and X. Cao, “Deep multi-scale context aware feature aggregation for curved scene text detection”, IEEE Trans. MM, 22, pp 1969-1984, 2020.

[8] Y. Liu, L. Jin and C. Fang, “Arbitrarily shaped scene text detection with a mask tightness text detector”, IEEE Trans. IP, 29, pp 2918-2930, 2020.

[9] S. Roy, P. Shivakumara, U. Pal, T. Lu, G. H. Kumar, “Delaunay triangulation based text detection from multi-view images of natural scene”, Pattern Recognition Letters, 129, pp 92-100, 2020.

[10] S. Nag, P. Shivakumara, U. Pal, T. Lu and M. Blumenstein, “A new unified method for detecting text from marathon runners and sports players in video”, Pattern Recognition, 107, 2020.

[11] M. Xue, P. Shivakumara, C. Zhang, Y. Xiao, T. Lu, U. Pal and D. Lopresti, “Arbitrarily-oriented text detection in low light natural scene images”, IEEE Trans MM, 2020.

[12] V. Carbune, P. Gonnet, T. Deselaers, H. A. Roweley, A. Daryin, M. Calvo, L. L. Wang, D. Keysers, S. Feuz and P. Gervais, “Fast multi-language LSTM-based online handwriting recognition”, IJDAR, 2020.

[13] C. Wang and C. L. Liu, “multi-branch guided attention network for irregular text recognition”, Neurocomputing, 2020.

[14] J. Zhang, C. Luo, L. Jin, T. Wang, Z. Li and W. Zhou, “SaHan: Scale-aware hierarchical attention network for scene text recognition”, Pattern Recognition Letters, pp 205-211, 2020.

[15] J. Lee, S. Park, J. Baek, S. J. Oh, S. Kim and H. Lee, “On recognizing text of arbitrary shapes with 2D self-attention”, In Proc. CVPRW, pp 2326-2335, 2020.

[16] S. Long, Y. Guan, K. Bian and C. Yao, “A new perspective for flexible feature gathering in scene text recognition via character pooling”, In Proc. ICASSP, pp 2458-2462, 2020.

[17] M. Shang, J. Gao and J. Sun, “Character region awareness network for scene text recognition”, In Proc. ICME, 2020.

[18] P. Verikas, A. Gelzinis, M. Bacauskiene, I. Olenin and E. Vaicukynas, “Phase congruency based detection of circular objects applied to analysis of phytoplankton images”, Patten Recognition, 45, pp 1659-1670, 2012.

[19] H. Chen, N. Xue, Y. Zhang, Q. Lu and G. S. Xia, “Robust visible infrared image matching by exploiting dominant edge”, Pattern Recognition Letters, 2018.

[20] X. Qin, X. Chu, C. Yuan and R. Wang, “Entropy-based feature extraction algorithm for stone carving character detection”, The Journal of Engineering, pp 1719-1723, 2018.

[21] S. Sharma, A. Sasi and A. Cheeran, “A SVM based character recognition system”, In Proc. RTEICT, 2017.

[22] R Ghoshal, A Banerjee, “SVM and MLP Based Segmentation and Recognition of Text from Scene Images Through an Effective Binarization Scheme”, In Proc. CIPR, pp 237-246, 2019.