ICIC Express Letters
Volume **17**, Number **11**, November **2023**
 ISSN 1881-803X
pp. **1277–1285**

A ROBUST TEXT DETECTION ON CERTIFICATE DOCUMENT USING PRUNED MAXIMUM TREE

Ednawati Rainarli[1,3], Suprapto[2,*] and Wahyono[2]

[1]Doctoral Program in Computer Science, Department of Computer Science and Electronics
[2]Department of Computer Science and Electronics
Faculty of Mathematics and Computer Sciences
Universitas Gadjah Mada
North Sekip, Bulak Sumur, Yogyakarta 55281, Indonesia
wahyo@ugm.ac.id
*Corresponding author: sprapto@ugm.ac.id

[3]Department of Informatics Engineering
Faculty of Engineering and Computer Science
Universitas Komputer Indonesia
Dipatiukur 114-116, Bandung 40132, Indonesia
ednawati.rainarli@email.unikom.ac.id

Received November 2022; accepted February 2023

Abstract. *This study discusses text detection on the certificate document image. Non-uniform font and spacing, various backgrounds of certificates, and low-resolution images are the challenges in detecting document images. We propose a detection model using the Pruned Maximum Tree. This method generated the text candidates from grayscale, negative grayscale, saturation image, and negative saturation. The process continued with geometric filtering and pruning of the branch of the tree. Two-stage classification is applied to creating the text classifier. In the initial stage, we carry out text classification and then classify ambiguous text candidates. The text classification implemented the Hough Orientation Gradient as features. The performance of the Pruned Maximum Tree was 10% better than the Maximum Stable Extreme Region. The most striking result is an increase of 20% in the recall. With the TedEval protocol, the F-measure of text detection obtained 76.82%. The improvement occurred when the detection applied the Support Vector Machine for text classification.*

Keywords: Certificate image, Document image detection, Maximum trees, Support vector machines, TedEval

1. **Introduction.** The rapid development of technology impacts the need for document digitization. We can convert a copy of a printed document to digital by using a scanner, such as a digital camera or a cell phone camera [1]. Retrieval of information through image documents is research that supports data processing in information systems. The trigger is the ease of use of the camera phone. It makes the input data more flexible [2-4]. The success of the extraction framework starts from the sensibility of the text detection [5]. Documents such as brochures, book covers, magazines, or certificates have more complex background images, variations in size, and different types of letters when compared to text in books, letters, or newspapers [6]. This condition becomes a challenge when detecting text in document images such as certificates. In addition, using images from mobile phone cameras is an obstacle in the detection process. The presence of low-resolution [7], sensor out of focus, and shadow when taking pictures create difficulties [2]. These reasons make the Optical Character Recognition (OCR) approach to document images with white background images unable to work optimally [8].

DOI: 10.24507/icicel.17.11.1277

Text detection can use a bottom-up approach, where the detection is initiated by generating the character components and then gradually combining them into text [9]. One of the bottom-up methods is the Maximally Stable Extremal Region (MSER) [10]. The method segments the candidate text as white connected areas. The segmentation applies some threshold values that increase gradually from 0 to 255. Images with area consistency and meeting the neighbouring rules of connected components are candidates for text components. This method is resistant to changes in scale and orientation. Detection text in document images [11] or natural images [12,13] used the MSER to generate text candidates. The results [12,13] show that the recall value of the detection using the MSER is lower than the precision value. One of the reasons is that the candidate text failed to be detected by the MSER. Not all text in images can meet the definition of extreme region [14]. Based on this argument, our research used the Maximum Component Tree (Maximum Tree) to generate candidate text components. The tree is a generalization of the MSER with the result that the Maximum Tree can produce a better recall than the MSER. However, using the Maximum Tree has a challenge [14]. The challenge is that the Maximum Tree generates more text candidates than MSER. This fact makes it difficult for us to filter the text components. Therefore, a strategy is needed to reduce non-text candidates. We propose pruning the maximum tree to filter candidate text so it can produce better text detection than using the MSER.

The other problem is the failure of the classifier to detect text candidates. Sometimes some of the generated text candidates appear as text partly or only one letter from the word it should be. The training process will fail to group the candidates into text. Therefore, we used a two-stage training process to sort out candidate texts and non-texts. The first applies the classification to forming text and non-text models. The second classification is to get a classifier for detecting ambiguous text. Main contributions to this study are as follows: 1) This study presents the automatic text detection model for certificates; 2) We introduce the Maximum Tree truncation rule for generating candidate text; 3) The two-stage training was applied to overcoming ambiguous text components. The structure of this paper is as follows. We start with the background of the problem in Section 1. We continue to discuss related studies on text detection and the Pruned Maximum Tree (PMT) in Section 2. Section 3 discusses the proposed detection method. Section 4 presents the results of the detection and analysis tests. Finally, we conclude the experiment result related to the certificate detection analysis in Section 5. This section also discusses future work for certificate detection research.

2. **Related Works.**

This section describes the development of text detection using the camera. In addition, it describes the development of scene text detection research using connected components, particularly MSER.

2.1. **Camera-based text detection studies.**

Taking images using a camera poses challenges in detecting text. [5] focuses on text detection in blurry image conditions and uses the zoning feature to classify text. The proposed features can recognize multilingual text, but the static block division has limitations. It makes smaller or larger the detected text block than the actual text size. In contrast to [15], this study uses the region growing for text detection on billboards. The research took an image from digital cameras. The detection is still limited to detecting text with a homogeneous background image. The result is like Khan and Mollah's research [16]. This study used Otsu segmentation to isolate and separate foreground pixels from background pixels. The proposed method can detect different sizes of text, but the test results use images with homogeneous background images. This condition is not like the text on the certificate. It is possible that the background image of the certificate is not homogeneous and has a complex background image. For this reason, several studies have adopted text detection techniques in scene images [12-14].

2.2. **Extraction with MSER and the Maximum Tree.** The review from [4] describes the grouping of text detection on natural images based on the technique used. One approach to detecting text is region-based. The MSER is widely used to detect text in scene images [12,13,17]. Neumann and Matas [10] initiated the MSER for detecting text in natural images. Over time there have been several improvements to MSER. [12] added a Fast-Guided Filter to overcome the appearance of low-resolution text and blurred images. Zhang et al. [13] added color information to MSER to detect the presence of text characters. Sun et al. [14] proposed a maximum tree approach to improve recall detection. This study used the same rules to remove non-text candidates as [14].

Based on the literature review, it can be concluded that region growing, or Otsu segmentation can only separate a homogeneous background color into two regions. However, this method is not sufficient for detecting text on certificates, as certificates often have a variety of background images, font types, sizes, and text colors. MSER segmentation, which is commonly used for detecting text in natural images, may be an alternative for detecting text in certificate images. However, the detection results with MSER have a low recall value. To address this issue, we propose using a generalized version of MSER, known as the Maximum Tree approach. One challenge of this approach is reducing the number of repeated text candidates that appear. To address this challenge, we propose a rule to remove repeated candidates from the Maximum Tree.

3. **Proposed Method.** There are five main steps to detect text in the certificate image, namely image conversion, generation of connected component candidates, filtering, classification, and removing the overlapping bounding boxes. Figure 1 shows the stages of the detection process. The preprocessing step involves converting the image to the HSV (Hue, Saturation, and Value) and creating four versions of the image: gray, negative gray, saturation, and negative saturation. This conversion to the HSV color space helps to separate the image's luminance (brightness) from its color information, making the text detection more resistant to changes in external lighting. The Maximum Tree method generates text candidates. According to [18], the Maximum Tree (Max-Tree) is a morphological data structure that describes the relationship between connected components of different threshold values. The Maximum Tree, which is a generalization of MSER, has the advantage of being resistant to changes in text size and orientation. Even when there is a lot of noise in the image, MSER can effectively identify stable regions due to its robustness to noise.

The implementation of the Maximum Tree algorithm refers to [19]. Component candidates are selected in the filtering process using geometric rules. The rules are the area of the connected component, height, weight, and filling rate. Equation (1) calculates the value of the variation area between the connected components and their parent.

$$r_i = \frac{A_{p_i} - A_i}{A_i} \tag{1}$$

where r_i is the variation area of the ith connected component, A_{p_i} is the parent area of the ith connected component, and A_i is the area of the ith connected component.

Each parameter followed the threshold value of the previous study [20]. The process continues with cutting the text candidates. Examples of component candidates are shown in Table 1. The first column is the index of the connected components, and the last column is the index of the parent. In Table 1, the leaves of the Maximum Tree are components with index numbers 17, 33, and 42. The branches are 19, 34, and 44. The pruning tree takes the nearest node of the branch tree. From Table 1, the pruning takes the component with index numbers 19, 34, and 44 as text candidates. This process removes the repeated candidates.

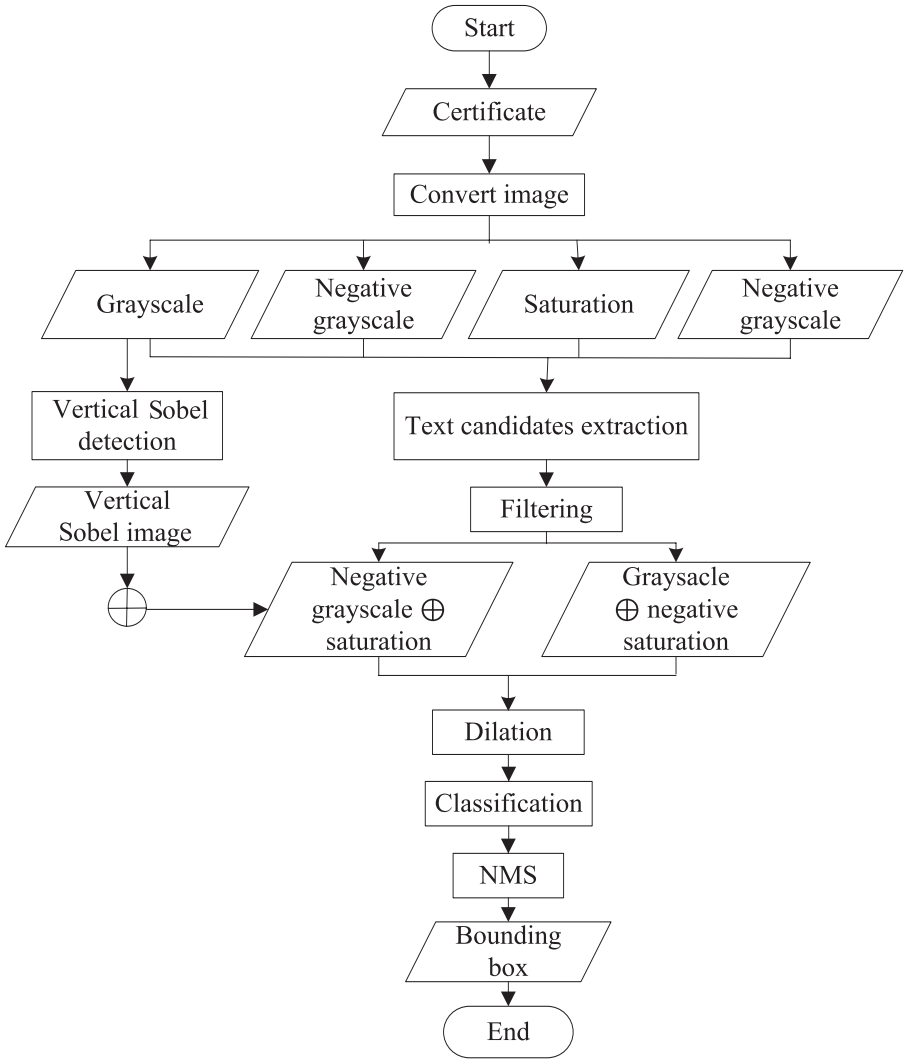

FIGURE 1. Certificate image detection process

TABLE 1. Example of connected components properties

Index	X_{min}	X_{max}	Y_{min}	Y_{max}	Area	Parent
17	818	833	999	1001	20	18
18	818	834	999	1001	24	19
19	817	835	999	1001	27	20
33	773	785	999	1001	20	34
34	773	786	999	1001	22	35
44	751	769	999	1001	25	45

After pruning, the process combines the filtered image with the vertical Sobel, followed by dilation. The dilation aims to connect candidate texts to form a word. Figure 2(a) shows the initial image, and Figure 2(b) shows an example of a negative grey image. Figure 2(c) shows the detection results using the Maximum Tree. In Figure 2(d), we applied the vertical Sobel. After the dilation process from Figure 2(e), the rest of the bounding box is classified using a machine learning algorithm. The classification removes the repeated bounding boxes. Non-Maximum Suppression (NMS) is a final step to removing the redundant bounding box. Finally, Figure 2(f) shows the result of detected bounding boxes.

FIGURE 2. Examples of detection results: (a) Initial image, (b) negative grayscale image, (c) the Maximum Tree result, (d) the vertical Sobel, (e) text candidates, and (f) final bonding box

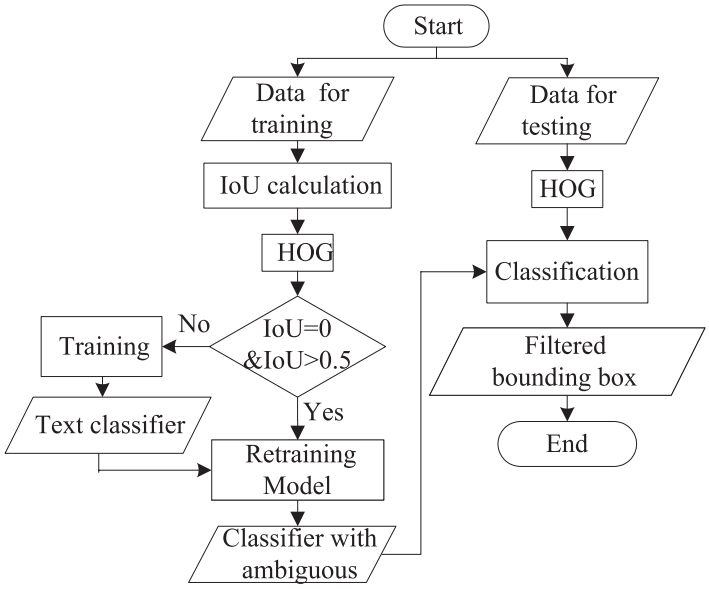

FIGURE 3. Stages of classification

There are two training processes in the classification. The first training used a text component with an Intersection over Union (IoU) value of more than 0.5 and a non-text with an IoU value equal to 0. The process adds those components into text and non-text training data, and then retrains the classification model. We extracted the Histogram of Gradient (HOG) features and then trained the model. This model classifies the data with IoU values between 0 and 0.5. Figure 3 shows the details of the classification. For HOG extraction, we resize each candidate text into 64×16. The size of each cell of the HOG is 4×4, and the number of cells per block is 2×2 with nine orientations. The HOG extraction produces 1620 features for each candidate text component.

To measure the success of text detection on certificates, we use the TedEval protocol [21]. In contrast to the DetEval protocol [22], the TedEval measurement has the advantage that it can detect the success of letter detection without having to annotate ground truth at the character level. We compare the text of the NMS results with the ground truth of each certificate test data. The filtered bounding box is continued with removing the repeated bounding boxes.

4. **Result and Discussion.** This study used 100 certificate data which we annotated at the word level. The annotation uses a web annotator to annotate the word text [23]. The annotation is in the form of a rectangle with YOLO Version 4 format. After the conversion, we get four bounding box coordinate values for each word from each certificate. We obtained 6203 words and then split 70 certificate data for training and model validation. The rest of the data is applied to testing the trained model. Several algorithms are implemented for recognizing text and non-text groups. Some of these algorithms are Extreme Learning Boosting (XGB), Random Forest (RF), Support Vector Machine (SVM), and Naïve Bayes (NB).

Table 2 represents the test results of the four algorithms after carrying out the training process from Figure 3. We observe the magnitude of the effect of using the Intersection over Union (IoU) value on the performance of each algorithm. The SVM is the algorithm with the highest performance when we compare it with RF, NB, and XGB. This result is consistent with several studies using SVM to classify text and non-text [10,24]. In addition, XGB also has a competitive performance compared to SVM. The challenge of XGB is that needing adjusted many hyperparameter values. Among these four algorithms, the NB had the lowest performance. We suspect this is because the data do not have linear characteristics. We got the same result when using linear kernel functions on SVM.

Table 2. Comparison of performance of algorithms with IoU values

Algorithm	TIoU = 0.3			TIoU = 0.4			TIoU = 0.5			TIoU = 0.6		
	P	R	F	P	R	F	P	R	F	P	R	F
XGB	0.96	0.90	0.93	0.97	0.92	0.94	0.97	0.93	0.95	0.97	0.95	0.96
RF	0.91	0.88	0.90	0.90	0.93	0.91	0.95	0.91	0.93	0.96	0.92	0.94
SVM	0.96	0.92	**0.94**	0.93	0.95	**0.94**	0.98	0.95	**0.97**	0.98	0.96	**0.97**
NB	0.67	0.91	0.77	0.66	0.93	0.77	0.65	0.95	0.77	0.63	0.95	0.76

TIoU = threshold value of IoU, P = precision, R = recall, F = F-measure

Table 2 explains that increasing the IoU value improves the performance of a classifier. However, increasing the accuracy will make the classification model have limitations on the ability to detect text. The reason is that the increasing IoU value causes reduced text candidates. This effect is seen in Table 3. Table 3 represents a result from testing data using the TedEval protocol. Even if IoU > 0.6 achieves the highest precision, the recall value will decrease when we increase the IoU value. Therefore, our next test uses IoU > 0.5.

Table 3. Text detection results on changes in IoU values

	Precision				Recall				F-measure			
IoU threshold	0.3	0.4	0.5	0.6	0.3	0.4	0.5	0.6	0.3	0.4	0.5	0.6
XGB	0.75	0.79	0.80	0.80	**0.75**	0.73	0.72	0.72	0.75	0.76	0.76	0.76
RF	0.73	0.72	0.78	**0.81**	0.73	**0.75**	0.71	0.70	0.73	0.73	0.75	0.75
SVM	0.76	0.77	0.79	**0.81**	0.73	**0.75**	0.74	0.74	0.74	0.76	**0.77**	0.76
NB	0.51	0.52	0.54	0.54	**0.75**	**0.75**	**0.75**	**0.75**	0.61	0.62	0.62	0.62

The further test, Table 4, compares the F-measure of one-stage and two-stage training. The first training classifies candidate text and non-text. Our first training model is used to sort out ambiguous text groups ($0 < \text{IoU} \leq 0.5$). There are positive and negative effects of using the two-stage training. The positive side is that the classification works effectively in the initial training process. The phenome means that the classifier model formed in the first training has a high F-measure. In Table 4, the XGB and SVM are the algorithms that work best for this case. The reason is that the F-measure of the two algorithms increases in the second training. This condition contradicts the RF and NB algorithms. The F-measure of the RF and NB tends to decrease. Different results are shown in MSER. Two-stage training sessions with ambiguous data are not better than using one-stage training. We suspect that two-stage training sessions will work optimally if the F-measure in the first training reaches 0.75.

Table 4. Comparison of F-measure for one-stage training with two-stage training

Algorithm	F-measure		Algorithm	F-measure	
	One-stage	Two-stage		One-stage	Two-stage
PMT, XGB	0.7593	0.7605	MSER, XGB	0.5153	0.5159
PMT, RF	0.7549	0.7473	MSER, RF	0.5234	0.5171
PMT, SVM	0.7660	0.7682	MSER, SVM	0.5227	0.5169
PMT, NB	0.6588	0.6240	MSER, NB	0.4112	0.3863

For final measurement, this research compared the classification result of Pruned Maximum Tree (PMT) with MSER. Table 5 presents no significant difference between PMT and MSER, except with NB. Both of accuracy and F-measure of these two methods are equally good in testing the classification model. We found notable difference when measuring text detection performance. The model has a better performance value with PMT than MSER. The measurement uses the TedEval protocol. In Table 6, the recall value increased to 20% in the SVM classification. Previous studies of scene text detection [12,25] also convey the problem of the low recall value of MSER. The bolded parts of Table 6 are the highest F-measure value of each PMT, MSER and EAST method. EAST stands for an Efficient and Accurate Scene Text detector. This method uses deep learning to detect text in natural images.

Table 5. Test results from classifier with PMT and MSER

Algorithm	Testing				Algorithm	Testing			
	Acc	P	R	F		Acc	P	R	F
PMT, XGB	0.99	0.98	0.96	0.97	MSER, XGB	0.99	0.98	0.98	0.98
PMT, RF	0.98	0.97	0.95	0.96	MSER, RF	0.98	0.97	0.96	0.96
PMT, SVM	0.99	0.99	0.97	0.98	MSER, SVM	0.99	0.98	0.98	0.98
PMT, NB	0.92	0.82	0.93	0.86	MSER, NB	0.83	0.76	0.90	0.78

Acc = accuracy, P = precision, R = recall, F = F-measure.

Table 6. Comparison of detection with PMT, MSER and EAST

Algorithm	P	R	F	Algorithm	P	R	F
PMT, XGB	0.8009	0.7239	0.7605	MSER, XGB	0.6614	0.4221	0.5153
PMT, RF	0.7845	0.7136	0.7473	MSER, RF	**0.7513**	**0.4016**	**0.5234**
PMT, SVM	**0.7941**	**0.7440**	**0.7682**	MSER, SVM	0.6760	0.4260	0.5227
PMT, NB	0.5357	0.7473	0.6240	MSER, NB	0.3691	0.4641	0.4112
EAST [26]	**0.8298**	**0.9088**	**0.8675**				

P = precision, R = recall, F = F-measure.

(a) (b)

FIGURE 4. Example of challenges: (a) Example of failed text that is detected as a bounding box; (b) example of several texts that is combined into one bounding box

In addition to presenting data on detection results using PMT and MSER, Table 6 also describes the performance of certificate detection using the EAST method [26]. We did not train on the EAST method but used a pre-train model to detect text in the certificate data. The SVM with the RBF function works best for text detection on the certificate. Detection limitations exist during the process of combining characters into the text. The dilation integrates candidate texts into words. The usage of dilation will experience problems if the distance between characters in the text is too large.

There are several challenges in detecting the text on the certificate image. The detection has difficulty when combining two closed lines. The problem causes the method to merge some text into one bounding box. If the text distance on the certificate is too far, the dilation fails to combine the characters into text words. The classification process deletes those bounding boxes. Figure 4(a) and Figure 4(b) describe examples of those problems. The "certificate" in Figure 4(a) and Figure 4(b) are not detected as one word. In Figure 4(b), the words "World Class Professor Program" are detected as one bounding box.

5. **Conclusions.** This research has proposed using PMT to extract text candidates. Test results show that PMT's extraction detects text on certificates better than MSER. This is supported by an increase in recall value of up to 24%. Two classification stages improve the performance results. However, this strategy has a negative impact if the classification method does not work optimally in the first training. The proposed detection model works best by using PMT and SVM. The certificate detection model effectively detects certificates with an F-measure of 76.82%. Future research plans to use other strategies to combine text components into words. Removing signatures that contact with text is also a concern to improve the success of text detection on certificates.

Acknowledgment. This work was supported by Directorate General of Higher Education, Research, and Technology, the Ministry of Education, Culture, Research, and Technology Indonesia under Grant No. 1891/UN1/DITLIT/Dit-Lit/PT.01.03/2022.

REFERENCES

[1] H. A. Niaz, A. Usman, M. U. Akram, A. Rafique and M. A. Khan, A study on word spotting techniques for document image analysis, *ACM Int. Conf. Proc. Series*, Taichung, Taiwan, pp.17-21, 2017.

[2] S. Lu, T. Chen, S. Tian, J. H. Lim and C. L. Tan, Scene text extraction based on edges and support vector regression, *Int. J. on Doc. Anal. and Recognit.*, vol.18, no.2, pp.125-135, 2015.

[3] S. Mahajan and R. Rani, Text detection and localization in scene images: A broad review, *Artif. Intell. Rev.*, vol.54, pp.4317-4377, 2021.
[4] E. Rainarli, Suprapto and Wahyono, A decade: Review of scene text detection methods, *Comput. Sci. Review*, vol.42, 100434, pp.1-24, 2021.
[5] S. A. Angadi and M. M. Kodabagi, A lightweight text extraction technique for hand-held device, *Int. J. of Image and Graphics*, vol.15, no.4, 2015.
[6] L. Wang, W. Fan, J. Sun, S. Naoi and T. Hiroshi, Text line extraction in document images, *The 13th Int. Conf. on Document Anal. and Recognit.*, Tunis, Tunisia, pp.191-195, 2015.
[7] X. Cao, W. Ren, W. Zuo, X. Guo and H. Foroosh, Scene text deblurring using text-specific multiscale dictionaries, *IEEE Trans. Image Process.*, vol.24, no.4, pp.1302-1314, 2015.
[8] A. P. Giotis, G. Sfikas, B. Gatos and C. Nikou, A survey of document image word spotting techniques, *Pattern Recognit.*, vol.68, pp.310-332, 2017.
[9] S. Eskenazi, P. Gomez-Krämer and J. M. Ogier, A comprehensive survey of mostly textual document segmentation algorithms since 2008, *Pattern Recognit.*, vol.64, pp.1-14, 2017.
[10] L. Neumann and J. Matas, A method for text localization and recognition in real-world images, *The 10th Asian Conf. on Comput. Vis.*, Queenstown, New Zealand, pp.770-783, 2010.
[11] K. Biswas et al., A new deep fuzzy based MSER model for multiple document images classification, *Pattern Recognit. and Artif. Intell.*, Paris, France, pp.358-370, 2022.
[12] R. Soni, B. Kumar and S. Chand, Text detection and localization in natural scene images using MSER and fast guided filter, *The 4th Int. Conf. on Image Inform. Process.*, Himacal Pradesh, India, pp.351-356, 2018.
[13] X. Zhang, X. Gao and C. Tian, Text detection in natural scene images based on color prior guided MSER, *Neurocomputing*, vol.307, pp.61-71, 2018.
[14] L. Sun, Q. Huo, W. Jia and K. Chen, A robust approach for text detection from natural scene images, *Pattern Recognit.*, vol.48, no.9, pp.2906-2920, 2015.
[15] R. M. Badiger, M. Y. Kammar and N. T. Pujar, Content extraction from advertisement display boards utilizing region growing algorithm, *2016 IEEE Int. Conf. on Advances in Electron., Commun. and Comput. Technol.*, Pune, India, pp.303-307, 2017.
[16] T. Khan and A. F. Mollah, A novel text localization scheme for camera captured document images, *Adv. in Intell. Syst. and Comput.*, vol.703, pp.253-264, 2018.
[17] A. Agrahari and R. Ghosh, Multi-oriented text detection in natural scene images based on the intersection of MSER with the locally binarized image, *Procedia Comput. Sci.*, vol.171, pp.322-330, 2020.
[18] R. Souza, L. Tavares, L. Rittner and R. Lotufo, An overview of max-tree principles, algorithms and applications, *The 29th Conf. on Graphics, Patterns and Images Tutorials*, Sao Paulo, Brazil, pp.15-23, 2017.
[19] L. Gueguen, *Maxtree*, https://github.com/gueguenster/maxtree, 2020.
[20] L. Sun and Q. Huo, An improved component tree based approach to user-intention guided text extraction from natural scene images, *The 12th Int. Conf. on Document Anal. and Recognit.*, Washington, D.C., United States, pp.383-387, 2013.
[21] C. Y. Lee, Y. Baek and H. Lee, TedEval: A fair evaluation metric for scene text detectors, *Int. Conf. on Document Anal. and Recognit. Workshops*, Sydney, Australia, pp.14-17, 2019.
[22] N. Nayef et al., ICDAR2017 robust reading challenge on multi-lingual scene text detection and script identification-RRC-MLT, *The 14th IAPR Int. Conf. on Document Anal. and Recognit.*, Kyoto, Japan, pp.1454-1459, 2017.
[23] Kili-Technology, *Annotation Tool for Images, Videos, Text*, https://kili-technology.com/annotation-tool/, 2021.
[24] L. M. Francis and N. Sreenath, TEDLESS – Text detection using least-square SVM from natural scene, *J. of King Saud University – Comput. and Inf. Sci.*, vol.32, no.3, pp.287-299, 2017.
[25] J. Dai, Z. Wang, X. Zhao and S. Shao, Scene text detection based on enhanced multi-channels MSER and a fast text grouping process, *The 3rd IEEE Int. Conf. on Cloud Comput. and Big Data Anal.*, Chengdu, China, pp.351-355, 2018.
[26] X. Zhou et al., EAST: An efficient and accurate scene text detector, *The 30th IEEE Conf. on Comput. Vis. and Pattern Recognit.*, Honolulu, United States, pp.2642-2651, 2017.