Cloud of Line Distribution for Arbitrary Text Detection in Scene/Video/License Plate Images

Wenhai Wang[1], Yirui Wu[1,2], Shivakumara Palaiahnakote[3], Tong Lu[1(✉)], and Jun Liu[4]

[1] National Key Lab for Novel Software Technology, Nanjing University, Nanjing, China
wangwenhai362@163.com, wuyirui@hhu.edu.cn, lutong@nju.edu.cn
[2] College of Computer and Information, Hohai University, Nanjing, China
[3] Department of Computer System and Information Technology, University of Malaya, Kuala Lumpur, Malaysia
shiva@um.edu.my
[4] School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, Singapore
jliu029@ntu.edu.sg

Abstract. Detecting arbitrary oriented text in scene and license plate images is challenging due to multiple adverse factors caused by images of diversified applications. This paper proposes a novel idea of extracting Cloud of Line Distribution (COLD) for the text candidates given by Extremal regions (ER). The features extracted by COLD are fed to Random forest to label character components. The character components are grouped according to probability distribution of nearest neighbor components. This results in text line. The proposed method is demonstrated on standard database of natural scene images, namely ICDAR 2015, video images, namely ICDAR 2015 and license plate databases. Experimental results and comparative study show that the proposed method outperforms the existing methods in terms of invariant to rotations, scripts and applications.

Keywords: Extremal regions · Text candidates · COLD
Text detection · License plate detection

1 Introduction

As technology grows, flexibility in capturing images and videos increases. This leads to have data of diversified images, such as scene images captured by high resolution camera, video images captured by low resolution camera and license plate images captured when vehicle is moving fast [1–3]. As results, one can

W. Wang and Y. Wu indicates equal contribution.

© Springer International Publishing AG, part of Springer Nature 2018
B. Zeng et al. (Eds.): PCM 2017, LNCS 10735, pp. 433–443, 2018.
https://doi.org/10.1007/978-3-319-77380-3_41

expect images affected by multiple diverse factors, such as low contrast, complex background, illumination effect, blur, occlusion etc. Therefore, there is tremendous scope for developing a method which can withstand the above multiple diverse factors. There are methods in literature, most methods focus on particular data with limited causes to achieve better results [1,2,4]. For example, Shivakumara et al. [5] proposed a new multi-modal approach to bib number/text detection and recognition in Marathon images. It combines biometric features with text feature to detect text in Marathon images. Recently, the methods are explored deep learning for text detection in natural scene images [6]. However, these methods require large number of labeled data and it is hard to optimize the parameters for different application and dataset [7]. Similarly, we can find many methods for text detection in video images. For example, Shivakumara et al. [8] proposed optical flow based dynamic curved video text detection. The method explores constant velocity and uniform magnitude properties of text components for text detection.

In the same way, there are methods for license plate detection in the images. In this work, we consider license plate detection is also text detection task because license plate is nothing but text with numerals. Moreover, it is challenging compared to scene and video data due to sever illumination effect, blur, contrast variations and distortions due to vehicle movements. Panahi and Gholampour [9] proposed accurate detection and recognition of dirty vehicle plate numbers for high speed applications. However, this method adapts specific language features for achieving results. In summary, it is noted that there are sophisticated methods for text detection in natural scene images, videos and license plate images. However, these methods focus on particular dataset with specific cause but not the images affected by multiple causes. Motivated by the work proposed in [10] for text spotting in natural scene images and license plate images, we propose a new method for text detection in all three types of images. The method used in [10] works well for high contrast images but not video images captured low resolution camera. Our aim is to address the challenges posed by natural scene images, video images as well as license plate images.

2 The Proposed Method

It is fact that text components in image of different types have characteristics which play a prominent role in representing character components, namely, shape, contrast, stroke information, uniform spacing between the characters. These characteristics are invariant to rotation, scaling, to some extent to distortions, illumination effects. To exploit such features, we propose Extremal Regions (ER) to detect text candidates which represent text. Due to background and foreground variations, ER couldn't detect text candidates accurately. Therefore, we propose filters to eliminate false text candidates. Then the proposed method uses polygonal approximation to detect dominant points on contour of the text candidates. The shape of the text candidates is studied by applying new concept called Cloud of Line Distribution (COLD) which finds unique property based

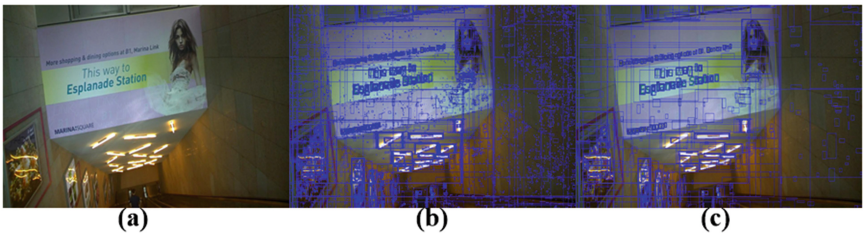

Fig. 1. Text candidates detection results, where (a) is the input image, (b) is the ER detection result and (c) represents the result of text candidate detection.

on angle information between dominant points. This results in character components. The character components are used to form text lines based alignment and spacing between character components.

2.1 Text Candidate Detection

For an input image, we generate text candidates, say $\{e_i\}$ of the input image I by detecting ER in multi-channels $\{C_l | l = 1...3\}$, namely, RGB channels [11]. Since the problem is complex, we propose the following filters for the output of ER.

(1) Filter using geometric properties: Since characters usually have similar geometric appearance, we estimate the ratio between the area and diameter of ER. In addition, this filter also used Euler number for eliminating false text candidates.
(2) Filter using intensity distribution: Inspired by the fact that characters have uniform color values. The proposed method discards the text candidates, which have high variation in intensity values. We first perform histogram operation on intensity values and adopt mean of the maximal and second-maximal value of the histogram as the split value. Then the proposed method calculates the intensity variance H_i of ER by

$$H_i = \frac{n_f \cdot \sum\limits_{x \in e_{i,f}} (I_x - A_{e_{i,t}})^2 + n_b \cdot \sum\limits_{x \in e_{i,b}} (I_x - A_{e_{i,b}})^2}{n_t + n_b} \quad (1)$$

where $e_{i,f}$ and $e_{i,b}$ represent the text and non-text of e_i respectively, n refers to the number of pixels and A represents the mean value. The effect of ER and Filtering is illustrated in Fig. 1 where (a) is the input image with arbitrary oriented text, (b) is the result of ER and (c) is the result of filtering. It is noticed from Fig. 1(c) that the above filtering classifies non-text candidates as text candidates. This is due to components in the background share filtering properties.

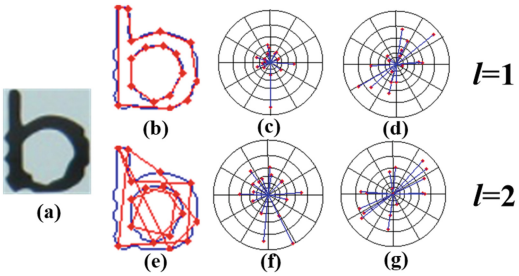

Fig. 2. Illustration of the process of the dominate points detection and COLD feature construction, where (a) is the example text ER, (b) and (e) define dominate point pairs with $l = 1$ and $l = 2$, (c), (f) and (d), (g) are the result distributions in Euclidean space and Tangent space, respectively.

2.2 Polygonal Approximation for Contour Points Detection

For each text candidate, the proposed method obtains Canny edges. The reason to choose Canny edge operator for edge contour detection is that Canny gives fine contours irrespective of contrast compared to Sobel and Prewitt which are sensitive to contrast variations. The dominant points p_k for each edge image of the text candidate are defined as

$$\begin{cases} \{c_j | j = 1...m\} = f_{cc}(Edge(e_i)) \\ \{p_{j,k} | j = 1...m, k = 1...n\} = f_{\varphi}(c_j) \end{cases} \tag{2}$$

where m and n refer to the number of connected components and dominant points, function $Egde()$ represents the Canny edge operator, function $f_{cc}()$ represents the operation of labeling connected components $\{c_j\}$ based on Canny edge image, function $f_{\varphi}()$ means the operation of detecting dominant points for each connected component c_j. The process of finding dominant points for the contour is known as polygonal approximation which is widely used in handwriting recognition [12] and shape classification [13]. Formally, we use the classical Ramer-Douglas-Peucker algorithm [14,15] as $f_{\varphi}()$ to detect dominant points, since it gives robust approximation results for complex contours and supports fast computing.

When we observe the output of polygonal approximation of the text candidates, text candidates representing actual characters have more straight lines while the text candidate representing false characters have more irregular lines. This observation motivates us to introduce shape descriptor called Cloud of Line Distribution (COLD) which extracts such observations in the form of distributions in angular space. This will be discussed in detail in subsequent section.

2.3 Cloud of Line Distribution for Character Component Detection

This section presents COLD feature extraction based on detected dominant points $\{p_{j,k}\}$, which describes shape of contours in multiple levels and space.

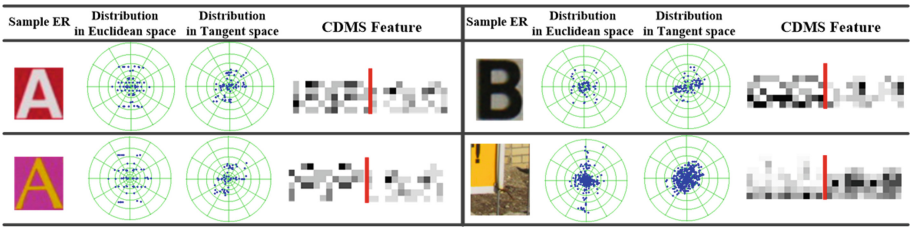

Fig. 3. The result cloud of point distributions and log-polar histograms based feature for different kinds of ER. Note that the red line is set to split feature into Euclidean-based and Tangent-based feature. (Color figure online)

Given an input image I and the ordered dominant points $\{p_{j,k}\}$, we aim to model the relation between pairs of dominant points in both Euclidean space $\mathcal{E}$ and Tangent space $\mathcal{T}$. Note that dominant points in $\mathcal{T}$ could be represented as $p_{\mathcal{T}} = (x, y, \rho)$, where we use Principle Components Analysis (PCA) to calculate tangent vector ρ and ρ is regarded as the orientation of this dominant point. In other words, ρ gives the direction based on neighbor information. We choose PCA and neighbor pixel information to find the orientation rather than choosing the gradient direction as used in the past for skeleton extraction [16] to focus on shape representation of contours and save the number of computations. The relation of dominant point pair $(p_{j,a}, p_{j,b})$ is thus measured as differences of orientation and distance, i.e. $\{\theta, d\}$, which could be represented as follows:

$$(p_{j,a}, p_{j,b}) = \begin{cases} d_{\mathcal{E}} = \sqrt{(x_b - x_a)^2 + (y_b - y_a)^2} \\ \theta_{\mathcal{E}} = \arctan(\frac{y_b - y_a}{x_b - x_a}) \\ d_{\mathcal{T}} = |\rho_a - \rho_b| \\ \theta_{\mathcal{T}} = f_{\varphi}(\rho_a - \rho_b) \end{cases} \tag{3}$$

where (x_a, y_a) and (x_b, y_b) are the coordinates of $p_{j,a}$ and $p_{j,b}$ respectively, ρ_a and ρ_b are the tangent vectors for $p_{j,a}$ and $p_{j,b}$ respectively and function $f_{\varphi}()$ computes the angle between a vector and horizontal line. Inspired by the Shape Context [17], we transform the $(\theta_{\mathcal{E}}, d_{\mathcal{E}})$ and $(\theta_{\mathcal{T}}, d_{\mathcal{T}})$ into log-polar space, which generates cloud of point distributions in multiple space. The two normalized log-polar histograms thus form the final feature vector $F_j(e_i) = [F_{j,e}, F_{j,t}]$ for the jth connected component, which is the proposed COLD feature and could be adopted as shape descriptors to classify text candidates as text or non-text. To describe relation between points in far distance, we define pairs every l points. In other words, we represent $p_{j,k}$ and $p_{j,k+l}$ as dominant point pair. Figure 2(b) and (e) show the process of defining pairs with $l = 1$ and $l = 2$ respectively and the corresponding result distributions with log-polar space are shown in Fig. 2(c), (f), (d) and (g). With variant values of l, we successfully capture properties of shape of contours in different levels. In other words, we describe shape of contours in local manner when l is small. When l is a larger value, we intend to construct a more global shape descriptors.

(a) (b) (c)

Fig. 4. The result of text line formation: (a) text candidates detection result, (b) text components after removing non-text components, (c) text line detection result.

We thus concatenate the COLD features with different l and different j together to form the final feature vector $F = [F_{j,k}(e_i)|j = 1...m, k = 1...l]$ for ith ER region. To present the description ability for shape of contours by our proposed COLD feature, we show the cloud of line distributions and log-polar histograms based feature F for different kinds of ER in Fig. 3. It is observed from Fig. 3 that the same character (Two A) tends to own the similar distributions and features, while different characters(A and B) or characters and background have different distributions and features. Finally, we apply COLD feature F in a random forest model to achieve the label for ith ER as $L = f_\tau(F(e_i))$. The final result of COLD is shown in Fig. 4, where (a) is text candidate image and (b) is the result of applying COLD. It is noted from Fig. 4(b) that all non-text candidates are removed.

2.4 Text Line Formation

Let text candidate be S which provides coarse locations of character components. To draw bounding box for the extracted text line, the proposed method groups the character components that share the common properties such as scale, orientation and color contrast. Inspired by [18], we find the color of character components which has uniform values compared to it background. Thus, we propose perceptual divergence to measure the perceptual divergence of a region against its surroundings as $PD(s) = \sum_{R,G,B} \sum_{\delta=1}^{w} h_\delta(s) \log \frac{h_\delta(s)}{h_\delta(\tilde{s})}$, where the term the Kullback-Leibler divergence (KLD) measuring the dissimilarity of two probability distributions, $h_\delta(s)$ and $h_\delta(\tilde{s})$ represent the histograms of text candidate and its surroundings. In this way, the proposed method groups the text candidates into text lines with the following conditions:

$$\begin{cases} 2/3 \leq |H_p/H_q| \leq 3/2 \\ |\theta_p - \theta_q| \leq \pi/8 \\ |f_d(p,q) - (W_p + W_q)/2| \leq H_p + H_q \\ |PD(p) - PD(q)| \leq 7 \end{cases} \tag{4}$$

where H, W and θ represent the height, width and orientation of text region respectively, and $f_d(p,q)$ refers to the distance between center of p and q.

Note that all the parameters in Eq. (4) are determined experimentally. If character components satisfy this property, the proposed method considers the character components for grouping as shown in Fig. 4(c), where one can see how the proposed method fixes bounding boxes for different oriented text lines.

3 Experiments

To evaluate the robustness of the proposed method, we consider three benchmark databases, namely, ICDAR 2015 scene [24], ICDAR2015 video [24] and Medialab License Plate database [22]. We select 10 videos from ICDAR2015 video dataset for our experiments. Medialab LPR is designed for license plate recognition and complex due to the images affected by severe illumination, touching characters, blur effect and perspective distortion effect. The evaluation scheme and instructions given in [24] are used for calculating standard measures, namely, Recall, Precision, F-measure and time cost per image. Note that the results of Table 1 are directly sampling from website of ICDAR 2015 scene competition [25]. In order to show effectiveness of the proposed method for video images, we implement Wu et al. [20] which explore character shape restoration for text detection in both natural and video images, Huang et al. [21] which uses the concepts of MSER and convolutional neural networks for text detection in natural scene images. In the same way, we use available code of the Yin et al. [19] and Neumann and Mattas [11] which use MSER concept for text detection in natural scene images. We also implement Anagnostopoulos et al. [22], Zhu et al. [23] and Zambeerletti et al. [10] to compare the effectiveness for text detection on licence plate images.

Table 1. Performance of text detection on ICDAR 2015 scene

Method	Precision	Recall	F-measure	Time-cost
Proposed	0.72	0.55	0.62	3.44
Only Euclidean space	0.70	0.51	0.59	3.42
Only Tangent space	0.69	0.52	0.59	3.43
CTPN	**0.74**	0.52	0.61	**1.40**
CNN Pro	0.35	0.34	0.35	–
HUST	0.44	0.38	0.41	–
NJU-Text	0.70	0.36	0.47	–
MSRA-v1	0.74	0.85	**0.79**	–

Quantitative results of the proposed and existing method are reported in Tables 1, 2 and 3 for the ICDAR 2015 scene data, ICDAR 2015 video data and License Plate data, respectively. Note that we achieve our results on a Laptop with 2.2 GHz Core2 i7, 6 GB RAM and Nvidia GTX 960M. Table 1 shows that

Table 2. Performance of text detection on subset of ICDAR 2015 video

Method	Precision	Recall	F-measure	Time-cost
Proposed	0.73	**0.65**	**0.69**	1.74
Only Euclidean space	0.72	0.60	0.65	1.72
Only Tangent space	0.70	0.62	0.66	1.72
Yin et al. [19]	0.64	0.57	0.60	**1.54**
Neumann et al. [11]	0.48	0.58	0.53	1.78
Wu et al. [20]	0.51	0.61	0.55	1.77
Huang et al. [21]	**0.79**	0.60	0.68	3.91

Table 3. Performance of text detection on Medialab LPR

Method	Precision	Recall	F-measure	Time-cost
Proposed	**0.90**	**0.81**	**0.85**	1.23
Only Euclidean space	0.85	0.78	0.81	1.02
Only Tangent space	0.85	0.73	0.78	1.03
Anagnostopoulos et al. [22]	0.81	0.63	0.71	0.87
Zhu et al. [23]	0.82	0.73	0.77	0.70
Zambeerletti et al. [10]	0.83	0.76	0.79	**0.61**

the proposed method is the second best at F-measure and Recall and third best at Precision for the ICDAR 2015 scene database. Focusing on shape information of text, the proposed method even outperforms MSER+CNN method, i.e. CNN Pro and NJUText, in f-measure. These facts show the importance of shape information for text detection task. Meanwhile, the proposed method would fail with cases where text appear with broken shape. Considering that most of the characters in ICDAR 2015 scene own solid shape, the proposed method thus achieves a good precision and recall performance. Several CNN-based methods, i.e. CTPN and MSRA-v1, score higher than our method in ICDAR 2015 scene. However, the proposed method achieves detection results with much lower time, computation and hardware cost. From Table 2, the proposed method scores the best results at Recall and F-measure and second best at Precision for subset of ICDAR 2015 video database compared to existing methods. Based on detection results of ICDAR 2015 scene and video, we could conclude the proposed method remain consistent in performance for various scenarios. Table 3 shows the proposed method score best in precision, recognition and F-measure but the last in time cost. The listed three methods are designed especially for car plate detection task. In that case, they try to get balance between computation cost and performance. Meanwhile, our method could handle with various situations, such as scene, video and car plate text. Qualitative results of the proposed method for the different database are shown in Fig. 5 where we can notice that the proposed method detects text well regardless of database, orientation, background

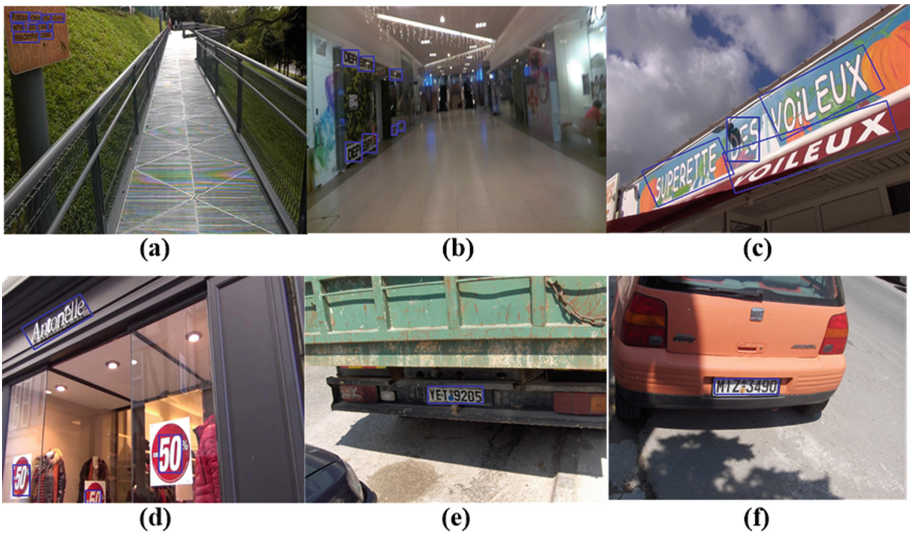

Fig. 5. Detection examples of the proposed method on ICDAR 2015 scene, ICDAR 2015 video and Medialab LPR.

complexity, font size and font variations. In summary, we can conclude that the proposed method is robust and generic as it gives consistent results for the different databases and outperforms the existing methods.

4 Conclusion

We have proposed a new method for arbitrary text detection in natural scene, video and license plate images. The proposed method explores ER for detecting text candidates. For text candidates, the proposed method uses polygonal approximation for dominant point detection over contour of text candidates. To eliminate false text candidates, we have introduced Cloud of Line Distribution which results in character components. Experimental results show the proposed method outperforms the existing methods for three different databases.

Acknowledgements. This work was supported by the Natural Science Foundation of China under Grant 61672273, Grant 61272218, and Grant 61321491, by the Science Foundation for Distinguished Young Scholars of Jiangsu under Grant BK20160021, by the Science Foundation of Jiangsu under Grant BK20170892, by the Fundamental Research Funds for the Central Universities under Grant 2013/B16020141 and by the open Project of the National Key Lab for Novel Software Technology in NJU under Grant KFKT2017B05.

References

1. Ye, Q., Doermann, D.S.: Text detection and recognition in imagery: a survey. IEEE Trans. PAMI **37**(7), 1480–1500 (2015)
2. Yin, X., Zuo, Z., Tian, S., Liu, C.: Text detection, tracking and recognition in video: a comprehensive survey. IEEE Trans. IP **25**(6), 2752–2773 (2016)
3. Weng, Y., Shivakumara, P., Lu, T., Meng, L.K., Woon, H.H.: A new multi-spectral fusion method for degraded video text frame enhancement. In: Ho, Y.-S., Sang, J., Ro, Y.M., Kim, J., Wu, F. (eds.) PCM 2015, Part I. LNCS, vol. 9314, pp. 495–506. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24075-6_48
4. Roy, S., Shivakumara, P., Mondal, P., Raghavendra, R., Pal, U., Lu, T.: A new multi-modal technique for bib number/text detection in natural images. In: Ho, Y.-S., Sang, J., Ro, Y.M., Kim, J., Wu, F. (eds.) PCM 2015, Part I. LNCS, vol. 9314, pp. 483–494. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24075-6_47
5. Shivakumara, P., Raghavendra, R., Qin, L., Raja, K.B., Lu, T., Pal, U.: A new multi-modal approach to bib number/text detection and recognition in marathon images. Pattern Recogn. **61**, 479–491 (2017)
6. Jaderberg, M., Simonyan, K., Vedaldi, A., Zisserman, A.: Reading text in the wild with convolutional neural networks. IJCV **116**(1), 1–20 (2016)
7. Liu, W., Wang, Z., Liu, X., Zeng, N., Liu, Y., Alsaadi, F.E.: A survey of deep neural network architecture and their applications. Neurocomputing **234**, 11–26 (2017)
8. Shivakumara, P., Lubani, M., Wong, K., Lu, T.: Optical flow based dynamic curved video text detection. In: Proceedings of ICIP, pp. 1668–1672 (2014)
9. Panahi, R., Gholampour, I.: Accurate detection and recognition of dirty vehicle plate numbers for high-speed applications. IEEE Trans. Intell. Transp. Syst. **18**, 767–779 (2016)
10. Zamberletti, A., Gallo, I., Noce, L.: Augmented text character proposals and convolutional neural networks for text spotting from scene images. In: Proceedings of ACPR, pp. 196–200 (2015)
11. Neumann, L., Matas, J.: Real-time scene text localization and recognition. In: Proceedings of CVPR, pp. 3538–3545 (2012)
12. He, S., Schomaker, L.: Writer identification using curvature-free features. Pattern Recogn. **63**, 451–464 (2017)
13. Wang, X., Feng, B., Bai, X., Liu, W., Latecki, L.J.: Bag of contour fragments for robust shape classification. Pattern Recogn. **47**(6), 2116–2125 (2014)
14. Ramer, U.: An iterative procedure for the polygonal approximation of plane curves. Comput. Graph. Image Process. **1**(3), 244–256 (1972)
15. Prasad, D.K., Leung, M.K.H., Quek, C., Cho, S.: A novel framework for making dominant point detection methods non-parametric. Image Vis. Comput. **30**(11), 843–859 (2012)
16. Wu, Y., Shivakumara, P., Wei, W., Lu, T., Pal, U.: A new ring radius transform-based thinning method for multi-oriented video characters. IJDAR **18**(2), 137–151 (2015)
17. Belongie, S.J., Malik, J., Puzicha, J.: Shape matching and object recognition using shape contexts. IEEE Trans. PAMI **24**(4), 509–522 (2002)
18. Li, Y., Jia, W., Shen, C., van den Hengel, A.: Characterness: an indicator of text in the wild. IEEE Trans. IP **23**(4), 1666–1677 (2014)

19. Yin, X., Yin, X., Huang, K., Hao, H.: Robust text detection in natural scene images. IEEE Trans. PAMI **36**(5), 970–983 (2014)
20. Wu, Y., Shivakumara, P., Lu, T., Lim Tan, C., Blumenstein, M., Kumar, G.H.: Contour restoration of text components for recognition in video/scene images. IEEE Trans. IP **25**(12), 5622–5634 (2016)
21. Huang, W., Qiao, Y., Tang, X.: Robust scene text detection with convolution neural network induced MSER trees. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014, Part IV. LNCS, vol. 8692, pp. 497–511. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10593-2_33
22. Anagnostopoulos, C., Anagnostopoulos, I., Loumos, V., Kayafas, E.: A license plate-recognition algorithm for intelligent transportation system applications. IEEE Trans. Intell. Transp. Syst. **7**(3), 377–392 (2006)
23. Zhu, S., Dianat, S.A., Mestha, L.K.: End-to-end system of license plate localization and recognition. J. Electron. Imaging **24**(2), 023020 (2015)
24. Karatzas, D., Gomez-Bigorda, L., Nicolaou, A., Ghosh, S., Bagdanow, A., Iwamura, M., Matas, J., Neumann, L., Chandrsekha, V.R.: ICDAR 2015 competition on robust reading. In: Proceedings of ICDAR, pp. 1156–1160 (2015)
25. ICDAR 2015 robust reading competition. http://rrc.cvc.uab.es/?ch=4&com=evaluation&task=1&gtv=1