Text Detection in Natural Scene Image: A Survey

Shupeng Wang(✉), Chenglin Fu, and Qi Li

College of Communication and Information Engineering,
Xi'an University of Science and Technology, Xi'an 710054, China
l013366723@qq.com

Abstract. Text detection in natural scene image is the extraction of the text regions from a natural scene image. The extraction information can be used in the system of text recognition. The texts in natural scene image contain important information. Text detection is an important prerequisite for many computer vision applications, such as license plate recognitions system, information filtering system, automatic navigation and so on. Text detection as a real-life application has to quickly and successfully process the texts in different fonts and under different environmental conditions. It should also be generalized to process texts in different languages and directions. We categorize different text detection techniques according to the methods used for each stage, and compare them in terms of merits, demerits and performance. Feature forecasts of text detection in natural scene image are given at the end.

Keywords: Text detection · Natural scene image · Text information

1 Introduction

Along with the popularization of smart phone, tablets and smart wearable equipment, it is more and more convenient to capture high-quality scene images. The natural scene images contain wealth semantic information, such as road signs, posters, license plate and signboards. Text recognition from the natural scene image is the important prerequisite for many computer vision applications, such as automatic image understanding and content-based image analysis tasks. Efficient text detection is the foundation of semantic information extraction. To acquire the text regions in a natural scene image, a lot of techniques, for instance, object detection, image processing and pattern recognition, will be used. The variations of the images and text cause challenges in text detection.

The text detection system, which extracts the regions of text from a given natural scene image, can be composed of five stages. The first stage is to acquire the natural scene image using a camera. The view of the camera, the illumination and the image quantity should be considered. The second stage is to process the images for the follow-up work. The third stage is to extract the candidate character regions based on some text features. The fourth stage is to ensure the character regions using classifiers. The last stage is to extract the text regions. As the characters in natural scene image are

X.-L. Huang (Ed.): MLICOM 2016, LNICST 183, pp. 257–264, 2017.
DOI: 10.1007/978-3-319-52730-7_26

always exit in words. The characters are grouped into words based on some features. Figure 1 shows the structure of text detection process.

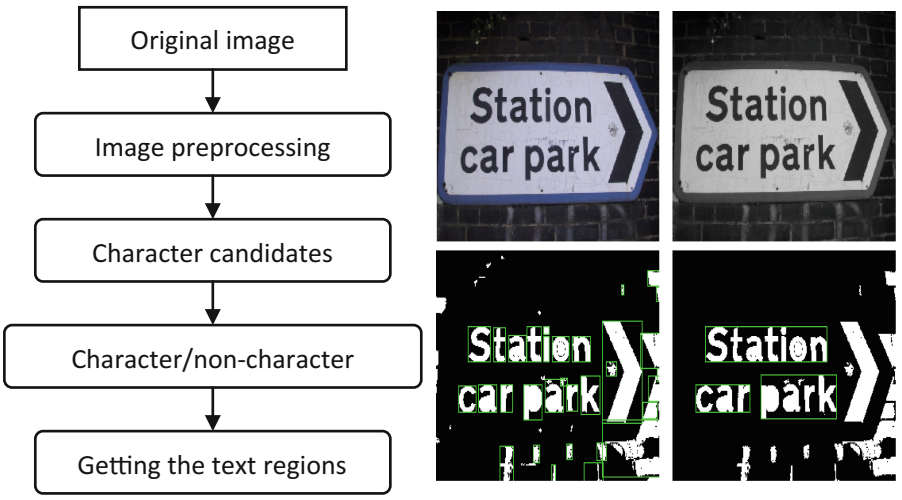

Fig. 1. On the left is the flowchart of the common text detection systems. On the right is the results of the main stages, including the original image, one of the channels after image preprocessing, the character regions and the text regions.

The reminder of this paper is organized as follows. In Sect. 2, the features of character are introduced. In Sect. 3, text extraction methods are classified with a detailed review. In Sect. 4, the protocols of the performance analysis of the Text Detection system are introduced. In Sect. 5, we summarize this paper and discuss areas for future research.

2 Character Features

The texts in images can be classified into two groups, the artificial texts and the scene texts. The artificial texts, such as the subtitle in the video, are artificially added. The scene texts, such as road signs and posters, are parts of the image. At the third stage of a text detection system, the character regions are extracted based on the features. The optional features are summarized as follows.

2.1 Morphological Features

(a) Color: the texts in natural scene image always exit as words, and the colors of the characters in the same words are always similar.

(b) Size: the characters in a same text line are always have the similar size. Though the size of the characters in an image may be various, the size may not be too big or too small.
(c) Distance: the characters in a same words may have similar distances, and the distances between adjacent words may have a certain ratio to the character distance.
(d) Shape: the character may have some specific shapes, such as aspect ratio and some holes. The area and perimeter can also be used.

2.2 Some Advanced Features

(e) Euler number: as the character regions may always have some holes, this feature means the difference between the number of connected components and the number of holes.
(f) Edge: the character regions may have difference with the background and wealth edge information.
(g) Horizontal crossings: the number of transitions between pixels belonging the region and not belonging to the region can computed to exclude the non-character regions.
(h) There are some more complex features, such as hole area ratio and convex hull ratio.

With the development of pattern recognition and digital image-processing, more and more features will be designed to improve the accuracy and efficiency of character regions detection.

3 The Techniques of Text Detection

According to the way of candidate character regions extraction, existing methods for scene text detection can roughly be categorized into two groups, texture-based methods and connected components-based methods.

3.1 Texture-Based Methods

The methods in this group exploit a sliding window and compute the texture features to search for the possible characters in the image. Then classifiers are used to identify the candidate character regions. At last group the regions into words.

In [2], the color clustering is used to extract the candidate character regions using the feature of pixel value. Then a support vector machine (SVM) is designed to remove the non-character regions. The adaptive mean shift algorithm (CAMSHIFT) is used to group the character regions into text regions. The method presented a detection rate of 87% using 50 images of different size, fonts and formats.

In [3], a special feature, combined the feature of HOG and the feature LBP, is designed to locate the characters in the image. Then cascade adaptive boosting (AdaBoost) classifier is adopted to ensure the character regions. To get the text regions, a window grouping method is used to generate text lines. At last a Markov Random Fields (MRF) model is used to filtered out the non-text regions. The method presented a recall rate of 67%, and the precision rate of 68% using the ICDAR 2003 Dateset.

In [4], Taking advantage of the desirable characteristic of gray-scale invariance of local binary patterns (LBP), a modified LBP operator is designed to extract the features of the characters. Then the classifier for is made by a polynomial neural network (PNN) to get the character regions. At last a post-processing procedure including verification and fusion is used to produce text regions. The method presented a recall rate of 87.7%, and the precision rate of 68% using the ICDAR 2003 Dateset.

In [5], six different classes features are used to extract the character regions. Then Modest AdaBoost with multi-scale sequential search is designed to get the text regions. The method use some complex features to improve the accuracy rate. However the complexity of the algorithm is also high. The method presented a recall rate of 75%, and the precision rate of 66% using the ICDAR 2003 Dateset.

In [6], AdaBoost is combined with Haar-like features to obtain cascade classifiers for text regions extraction. The method presented a recall rate of 79.9%, and the precision rate of 72.6% using 128 street view images.

Different from the above methods, in [7], wavelet transform is applied to the image and the distribution of high-frequency wavelet coefficients is considered to statistically characterize text and non-text areas. Then the k-means algorithm and projection analysis are used to detect and refine the text regions. The method presented a recall rate of 90%, and the precision rate of 87% using a set of video frames taken from the MPEG-7 video test set. Based on this methods, in [8], use a new Fourier-Statistical Features (FSF) in RGB space to detect texts of different fonts, size and scripts. The experimental results show that the method has made some improvement in terms of recall rate and precision rate.

The methods in this group use a sliding window to localize individual characters, or the whole words. Strengths of such method include robustness to noise and blur. The main drawback is how to define the size of the window. Since the too big windows may result in too much noise, and the too small windows may give rise to the difficulty of computing the features.

3.2 Connect Component Based Methods

The Connect component based methods extract candidate character regions by connect component. Then group the characters regions into text. And some additional checks may be used to refine the detection results.

As text in the natural scene images always have closely spaced edges, the edge feature can be used to detect the character regions. In [9], the Sobel operator is used to get edges. Then local thresholding and hysteresis edge recovery are applied to get the character regions. The projection analysis is used to group the character regions into

text regions. The method can process multilingual text characteristics, including English and Chinese.

In [10], the edges are detected by the wavelet transform and scanned into patches by a sliding window. Then a simple classification procedure with two learned discriminative dictionaries is applied to get candidate text areas. At last, adaptive run-length smoothing algorithm and projection analysis are used to refine the candidate text areas.

As text in the natural scene images always have special color, intensity and stroke width, these features can be used to detect some special connect component as the character regions. In [11], firstly the Stroke Width Transform (SWT) is applied to the image. Then detect the connect components with similar stroke widths as the candidate character regions. At last the text lines are built with the features of shape and distance. In [12], after getting the character regions by SWT, the color and shape of the regions are used to build multi-direction text lines.

After the concept of maximally stable extremal regions (MSER) is presented in [13], the feature region has been widely used in the system of image retrieval and object detection. In [1], the MSERs are detected as the candidate character regions. Then the connected component analysis (CCA) is used to get the text regions.

In [14], after detecting the MSERs, an efficiently pruned exhaustive search algorithm is used to filter out the nesting or duplicate regions. Then the morphological features and Single-link algorithm are used to group the character regions into text regions. The posterior probabilities of text candidates corresponding to non-text are estimated with a SVM classifier. And the Bayesian decision rule is used to refine the detection results. The method presented the best performance in ICDAR 2015 [15].

In [16], the extremal regions (ER), with a more simple calculation procedure, are detected as the candidate character regions. Then a two-stage algorithm is designed to pruning the non-character regions. In the first stage, SVM and five features are used to estimate the class-conditional probabilities of ERs. And in the second stage, some more complex features and AdaBoost classifier are used to refine the results. Finally, a clustering-based method is used to group the character regions into text regions.

The Connect component based methods, recently more popular approach, can detect most text regions in the natural scene images. And these methods are robustness to incline, rotation and blur. However, there are still drawbacks. One problem is that the number of the feature regions will be large. Another problem is the absence of an effective text candidates construction algorithm.

4 Evaluation Protocols

With so many approaches and datasets, reproducing all of them and comparing them with each dataset are problematic. For text detection, the ICDAR protocols are most commonly adopted. The ICDAR 03 [17] and ICDAR 11/13 [15] datasets are prepared for scene text, covering tasks of text location, character segmentation and word recognition. And the Street View Text (SVT), a more complex dataset, is another commonly used dataset.

In [18], Wolf and Jolion proposed the DteEval protocol that comprise the area overlap [17] and the object level evaluation. As shown in Fig. 2, it supports one-to-one, many-to-one and one-to-many matches among the ground truth and detections. This protocol was adopted in ICDAR 2011 and ICDAR 2013 "Robust Reading" competitions.

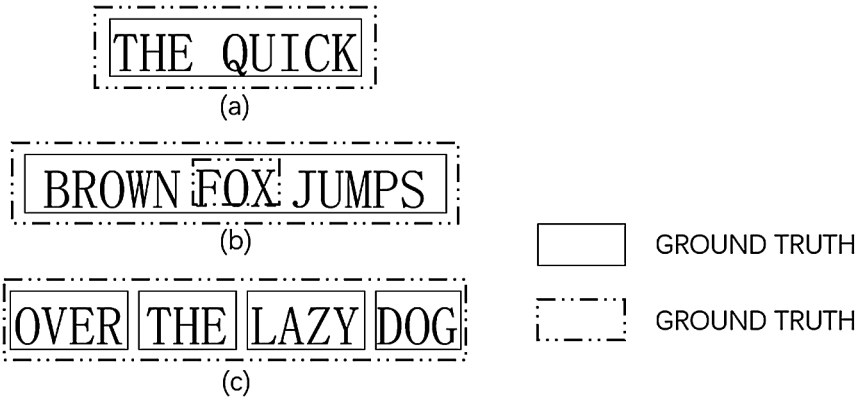

Fig. 2. Different match types between ground truth rectangles and detected rectangles: (a) one-to-one match; (b) a split: a one-to-many match with one ground truth rectangle; (c) a merge: a one-to-many match with one detected rectangle

The precision and the recall are used to measure the performance of the approaches. The precision is defined as the ratio between the area of intersection regions and that of detected text regions. Recall is defined as the ratio between the area of intersection regions and that of ground truth regions. In case of multiple images or a single image with multiple text rectangles, a natural way has been proposed to get the results. With saving the detection results and ground truth in XML format, there has been an available system to evaluation the algorithms.

$$Recall\ \ rate = \frac{N.o.correctly\ detected\ rectangles}{N.o.rectangles\ in\ the\ database} \tag{1}$$

$$Precision\ \ rate = \frac{N.o.correctly\ detected\ rectangles}{N.o.rectangles\ in\ the\ database} \tag{2}$$

5 Conclusion

This paper presented a comprehensive survey on existing techniques of text detection in natural scene images. Although significant process of text detection has been made in the last few decades, there is still a lot of work to be done since a robust system should work effectively under a variety of environmental conditions and text formats.

In most text detection systems, extracting the character regions and grouping them to text regions are always two independent stage. Taking advantage of the affiliation of characters and texts may provide some new methods for this problem. With the rapid expansion of machine learning technique, design more effective classifiers can be an important task in this filed.

References

1. Neumann, L., Matas, J.: A method for text localization and recognition in real-world images. In: Kimmel, R., Klette, R., Sugimoto, A. (eds.) ACCV 2010. LNCS, vol. 6494, pp. 770–783. Springer, Heidelberg (2011). doi:10.1007/978-3-642-19318-7_60
2. Kim, K.I., Jung, K., Jin, H.K.: Texture-based approach for text detection in images using support vector machines and continuously adaptive mean shift algorithm. IEEE Trans. Pattern Anal. Mach. Intell. **25**(12), 1631–1639 (2003)
3. Pan, Y.F., Hou, X., Liu, C.L.: A robust system to detect and localize texts in natural scene images. In: The Eighth IAPR International Workshop on Document Analysis Systems, pp. 35–42. IEEE (2008)
4. Ye, J., Huang, L.L., Hao, X.: Neural network based text detection in videos using local binary patterns. In: Chinese Conference on Pattern Recognition, CCPR 2009, pp. 1–5 (2009)
5. Lee, J., Lee, P.H., Lee, S.W., et al.: AdaBoost for text detection in natural scene. In: International Conference on Document Analysis and Recognition, pp. 429–434. IEEE Computer Society (2011)
6. Song, Y., He, Y., Li, Q., et al.: Reading text in street views using Adaboost: towards a system for searching target places. In: 2009 IEEE Intelligent Vehicles Symposium, pp. 227–232. IEEE (2009)
7. Gllavata, J., Ewerth, R., Freisleben, B.: Text detection in images based on unsupervised classification of high-frequency wavelet coefficients. Proc. Int. Conf. Pattern Recogn. **1**(3), 425–428 (2004)
8. Shivakumara, P., Phan, T.Q., Tan, C.L.: New fourier-statistical features in RGB space for video text detection. IEEE Trans. Circ. Syst. Video Technol. **20**(11), 1520–1532 (2010)
9. Lyu, M.R., Song, J., Cai, M.: A comprehensive method for multilingual video text detection, localization, and extraction. IEEE Trans. Circ. Syst. Video Technol. **15**(2), 243–255 (2005)
10. Zhao, M., Li, S., Kwok, J.: Text detection in images using sparse representation with discriminative dictionaries. Image Vis. Comput. **28**(12), 1590–1599 (2010)
11. Epshtein, B., Ofek, E., Wexler, Y.: Detecting text in natural scenes with stroke width transform. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition, pp. 2963–2970. IEEE (2010)
12. Yao, C., Bai, X., Liu, W., et al.: Detecting texts of arbitrary orientations in natural images. In: IEEE Conference on Computer Vision & Pattern Recognition, pp. 1083–1090 (2012)
13. Matas, J., Chum, O., Urban, M., et al.: Robust wide-baseline stereo from maximally stable extremal regions. Image Vis. Comput. **22**(10), 761–767 (2004)
14. Yin, X.C., Yin, X., Huang, K., et al.: Robust text detection in natural scene images. IEEE Trans. Pattern Anal. Mach. Intell. **36**(5), 970–983 (2013)
15. Karatzas, D., Gomezbigorda, L., Nicolaou, A., et al.: ICDAR 2015 Competition on Robust Reading. International Conference on Document Analysis and Recognition (2015)
16. Neumann, L., Matas, J.: Real-time lexicon-free scene text localization and recognition. 1 (2015)

17. Lucas, S.M., Panaretos, A., Sosa, L., et al.: ICDAR 2003 robust reading competitions. In: International Conference on Document Analysis and Recognition, p. 682. IEEE Computer Society (2003)
18. Wolf, C., Jolion, J.M.: Object count/area graphs for the evaluation of object detection and segmentation algorithms. Doc. Anal. Recogn. **8**(4), 280–296 (2006)