Pattern Recognition Letters 67 (2015) 153–162

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Detecting natural scenes text via auto image partition, two-stage grouping and two-layer classification[☆]

Anna Zhu, Guoyou Wang[*], Yangbo Dong

State Key Lab for Multispectral Information Processing Technology, Huazhong University of Science and Technology, Wuhan 430074, China

ARTICLE INFO

Article history:
Available online 18 June 2015

Keywords:
Text detection
Text region appearances
Image partition
Two-stage grouping
Two-layer classification

ABSTRACT

Text detection in natural scene images is important and challenging work for image analysis. In this paper, we present a robust system to detect natural scene text according to text region appearances. The framework includes three parts: auto image partition, two-stage grouping and two-layer classification. The first part partitions images into unconstrained sub-images through statistical distribution of sampling points. The designed two-stage grouping method performs grouping in each sub-image in first stage and connects different partitioned image regions in second stage to group connected components (CCs) to text regions. Then a two-layer classification mechanism is designed for classifying candidate text regions. The first layer is to compute the similarity score of region blocks and the second layer is a SVM classifier using HOG features. We add a normalization step to rectify perspective distortion before candidate text region classification which improves the accuracy and robustness of the final output result. The proposed system is evaluated on four types datasets including two ICDAR Robust Reading Competition datasets, a born-digital image dataset, a video image dataset and a perspective distortion image dataset. The experimental results demonstrate that our proposed framework outperforms state-of-the-art localization algorithms and is robust in dealing with multiple background outliers.

1. Introduction

Nowadays, more and more people use handhold devices equipped with high-resolution cameras to record scenes in their daily life. Information analysis of these scene images attracted much attention in computer vision field. Among them, text information is intuitive and plays a vital role for various applications such as content-based image retrieval, scene understanding and vision based navigation. Due to its immense potential for commercial applications and ability to be applied in human computer interaction field, research about text extraction is being pursued both in academia and industry. Generally, text extraction paradigms are decomposed to three parts: text detection, text binarization and text recognition (OCR). They solve the problems: “where is the text?”, “what is the text?” and “what is the text content?” respectively. In this paper, we focus on solving text detection problem. However, detecting text from natural scene images encounters many difficulties [10], for instance the clutter background, geometrical distortions, various text orientations and appearances. Much work in this area has been done, like yearly competitions [11,24] and new algorithms [10,31] to improve the accuracy and computational complexity issues.

Text detection algorithms can be roughly classified into two categories. One is to find text component groups. The other is to find text blocks. The former uses distinct geometric and/or color features based on CC level to find the candidate text (or character) components. This category is considered as content-based image partition grouping spatial pixels to connected components conditionally. Usually, geometrical analysis is applied to classify text component and non-text component. Remaining CCs with similar properties are gathered together to form text regions. This kind of methods may use local gradient features and uniform colors of text characters [30], maximally stable extremal region (MSER) algorithm [13], intensity histogram based filter and shape filter [17], stroke width transform [6], K-means clustering [26] and mathematical morphology based method for multilingual text detection [16]. The detected text components of this category can be used directly for text recognition. However, it will fail when text components are not homogeneous and the consumption time will increase with the complexity of the image since more CCs arise to be co-analyzed.

The alternative category attempts to classify regions in a given patch belonging to text regions or non-text regions by texture analysis and then merges neighboring text regions to generate text blocks. Usually, this method uses multi-scale strategy and partitions images

[☆] This paper has been recommended for acceptance by Sankar Kumar Pal.

[*] Corresponding author. Tel.: +86 186 2770 8137.
E-mail address: gywang@hust.edu.cn (G. Wang).

http://dx.doi.org/10.1016/j.patrec.2015.06.009
0167-8655/© 2015 Elsevier B.V. All rights reserved.

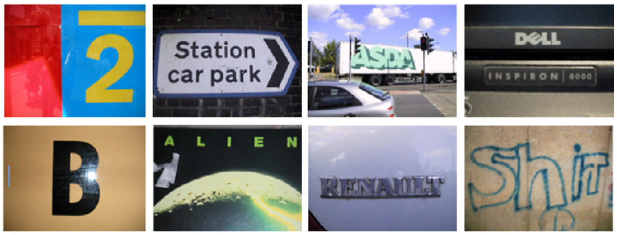

Fig. 1. Examples of different text appearances in natural scene images.

into patches with sliding windows. Then analysis the region feature like wavelet decomposition coefficients at different scales [29], gradient-based maps and histograms of block patterns [14] or different gradient edge features (mean, standard deviation, energy, entropy, inertia and local homogeneity) of image blocks [25] with different classification tools as SVM, AdaBoost or ANN. This kind of methods capture the inter-relationship of text and treat the character string as a whole which can detect texts accurately even when noisy. While the operating speed is relatively slow for its non-content based methods that regions are generated on dense sampled pixels or fixed step with different scales and sizes. Also, the performance will lose its advantage for non-horizontal aligned text.

To overcome the difficulties and take advantage of the above two categories, hybrid approaches are proposed. In Pan's paper [22], a region-based method is firstly used to estimate the text existing confidence and scale information in image pyramid. Then, a conditional random field (CRF) model is performed to filter out non-text components. While Huang [8] takes advantages of both MSERs and sliding window based methods by using convolutional neural network (CNN) to robustly identify text components from text-like outliers after MSERs operator which can dramatically reduce the number of windows scanned and enhances detection of the low-quality texts. These kind of methods first roughly select candidate text regions with one category and then use other strategies to confirm the confidence of text regions and filter out non-text regions.

In this paper, we use a different way to detect text depending on the appearance of detected candidate text regions. As shown in Fig. 1, natural scene texts present in different forms. They might be single character regions, connected character regions or individual character regions. Single character detection is more like character recognition so we do not consider this type. For the other two forms, we design a two layer filtering mechanism to classify text and non-text regions. Our proposed method belongs to CC-based method and uses the intrinsic characteristic of text and individual property of characters which can effectively detect text with different fonts, sizes, colors and invariant color and shape of attachment surface. The framework is a coarse-to-fine process starting from the whole to individual part then to integration. From the view of granulation [28], it involves the process of two directions: decomposition and construction. The decomposition involves the process of dividing a larger granule into smaller and lower level granules. It a top-down process. The construction involves the process of forming a larger and higher level granule with smaller and lower level sub-granules. It is a bottom-up process. Here text is considered as larger granule and CCs are lower level granules. We consider the text as an integral structure which is opposite to the MSER method. We then analysis the feature of each connected component in extracted sub-regions and group the candidate characters to the text string. Finally, a text candidate verification step is used to refine the detected text region result. The main contributions of our framework is listed as follows.

(1) Besides analyzing individual character features, text string structure can also be used for image partition. The rectilinear characteristic of text string structure and parallel edge pair characteristic of characters are employed to form location and gray level distribution of representative points. Based on the distribution, an image is partitioned to several sub-images. Each sub-image contains certain horizontal position and color information of potential text region.
(2) The proposed features for unary component classification are valid for both individual character components and jointed text string components. We also use two stage grouping method mixed with region-based analysis to locate candidate text regions which can detect multi-polarity text.
(3) In most previous CC-based methods, the grouping step only groups individual characters but ignores the fact that multiple characters may be jointed into a single connected component. In this paper, we propose the three types of text region appearances and analyze them individually.
(4) Characters in text strings are normally different and this particular characteristic could be utilized to compare the similarity of regions. The first layer of classification uses a similarity score based on this characteristic which can filter out most repeat backgrounds.

The rest of this paper is organized as follows. Section 2 details the proposed method. Experiments and results are presented in Section 3 and conclusions are drawn in Section 4.

2. The proposed method

In our method, we choose Y-Cr-Cb space and ab channels of L-a-b color space to find the text region for these channels show the best performance in our experimental result as shown in Section 3. Our method contains three parts: auto image partition, a two-stage grouping and a two-layer classification. The flowchart of the proposed approach is depicted in Fig. 2.

2.1. *Auto image partition*

In this stage, a natural scene image is partitioned to several sub-images without predefined parameters. We use a trained ANN classifier to classify CCs in each sub-image.

2.1.1. *Combined edge map from multi-channel*

As parallel edge is one of the most distinct features of text, firstly we extract edges in image. Unlike the normal Canny detection performed on single channel or gray scale level, we modify the Canny edge detector [2] to get an improved edge map with multi-channels. We use five channel images to convolve with a Gaussian filter sequentially and while calculating the intensity gradient and direction of each pixel in these images separately. Then, by searching each pixel's maximum intensity gradient among the five-channel images, we compose a single max-gradient map. Meanwhile, we record the corresponding direction in the orientation map. After that, non-maximum suppression is implemented on the max-gradient map in the gradient direction and two thresholds are set to find edges. In our

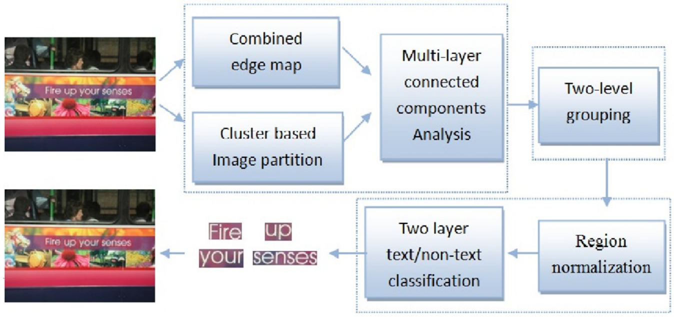

Fig. 2. The flowchart of our proposed method.