21st International Conference on Pattern Recognition (ICPR 2012)
November 11-15, 2012. Tsukuba, Japan

Scene Text Detection via Stroke Width

Yao Li and Huchuan Lu
School of Information and Communication Engineering, Dalian University of Technology, China
liyao.dut@gmail.com, lhchuan@dlut.edu.cn

Abstract

In this paper, we propose a novel text detection approach based on stroke width. Firstly, a unique contrast-enhanced Maximally Stable Extremal Region(MSER) algorithm is designed to extract character candidates. Secondly, simple geometric constrains are applied to remove non-text regions. Then by integrating stroke width generated from skeletons of those candidates, we reject remained false positives. Finally, MSERs are clustered into text regions. Experimental results on the ICDAR competition datasets demonstrate that our algorithm performs favorably against several state-of-the-art methods.

1. Introduction

In recent decades, detecting text in complex nature scenes is a hot topic in computer vision, since text in images provides much semantic information for human to understand the environment. Moreover, text detection is a prerequisite for a couple of purposes, such as content-based image analysis, image retrieval, etc. Unlike overlay text detection in video frames where lots of prior knowledge can be employed, text detection in natural scene images is a difficult problem due to complex background, variations in text's size, font, color, orientation and lighting conditions.

Generally, methods on this topic can be divided into two categories: learning-based methods and connected component (CC)-based methods.

In order to distinguish text regions from non-text ones, learning-based methods use some features to train a classifier (e.g., SVM or AdaBoost). Pan *et al.* [6] use a polynomial classifier in the verification stage and evaluate five widely used features, including HOG, LBP, DCT, Gabor filter and wavelet, then find the combination of HOG and wavelet showing the best performance. Wang *et al.* [9] use gray scale contrast feature and edge orientation histogram feature to train a SVM. The main limitations of learning-based methods are high computational complexity and the difficulty to select the best features for scene text detection.

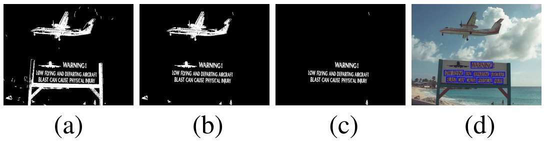


Figure 1. Overview of text detection process. (a) Detected MSERs. (b) CCs after geometric filtering. (c) CCs after stroke width extraction. (d) Detected text.

CC-based methods, on the other hand, usually generate separated CCs using some properties, such as edge, stroke width and color. After that, some geometric constraints are designed to remove false positives. Epshtein *et al.* [1] propose stroke width transform, which converts value of each color pixel into the width of most likely stroke. Zhang and Kasturi [11] use HOG to locate text edges and then Graph Spectrum is utilized to group the characters and remove false positives. The advantage of these methods is that their computational complexity is low. However, the performance of CC-based methods are likely to degrade when dealing with texts in complex background.

In this paper, a novel CC-based text detection algorithm is proposed to overcome the difficulties mentioned above. We make three major contributions compared with other methods available in literature. (1) Though MSER has been exploited in the text detection task, such as [5], most of those approaches use bare MSER algorithm, ignoring the fact that MSER is sensitive to image blur. We overcome this obstacle by incorporating intensity information on the boundary between text and background. (2) Since stroke width is one of the inherent properties of text, which is insensitive to size, font, color, orientation of text, stroke width on the skeleton of CCs is extracted to distinguish between text and non-text regions. (3) We only detect text on one scale, this is more efficient than the work [6] which requires image pyramid in order to detect text with different sizes.

978-4-9906441-1-6 ©2012 IAPR

2. Text Detection Algorithm

An overview of our algorithm is depicted in Figure 1. On every input color image, we first resize it into 640 × 480 (or 480 × 640) resolution, then MSERs are detected and considered as text region candidates (Section 2.1). As a next step, we design some simple heuristic rules to remove MSERs which are not text regions (Section 2.2). Different from stroke width transform in the work [1], we propose stroke width generated by distance transform on the skeleton of each CC to eliminate non-text areas (Section 2.3). In the final step, we group characters into words based on Euclidean distance, orientation and similarities between characters (Section 2.4).

2.1. Contrast-enhanced MSER Detection

The concept of MSER is introduced by Matas *et al.* [4]. Since a single letter usually shares similar color and its intensity is often quite different from background, MSER can locate these text regions efficiently. MSER has many good properties, such as invariance to affine transformation of image intensities, stability [4] etc., however, it is sensitive to image blur. An example demonstrating this is shown in Figure 3 (b). It is obvious that most of characters are blurred and connected, so it is really difficult for us to get true stroke width of every character in Section 2.3. In order to overcome this problem, we propose a novel contranst-enhanced MSER algorithm as follows.

For an input image I, based on the observation that there are large changes in intensity at the boundary between text pixels and background, an intensity image In is obtained as $In = (R + B + G)/3$ in HSI color space. After that, we check intensity gradient using $In(i+1, j) - In(i-1, j) > T_1$, where T_1 is a threshold, if this condition is met, then update:

$$I_C(i \pm 1, j) = I_C(i \pm 1, j) \pm T_2, \qquad (1)$$

where $C \in \{R, G, B\}$, parameter T_2 is a predefined threshold. The aim of this procedure is to enhance the contrast between characters and background (Figure 2). Finally, we conduct MSER detection on this contrast-enhanced image. Figure 3 (c) illustrates the result of our contrast-enhanced MSER detection where all letters in the same word are separated.

2.2. Geometric Filtering

After locating bounding boxes of MSER, we design some simple geometric rules to filter out obvious non-text regions. Firstly, by assuming all characters have been separated, we limit the aspect ratio of each bounding box between 0.3 and 3. Secondly, text region candidates with low saturation (less than 0.3) or small area (less than 30 pixels) are unlikely to be text regions, thus they should be removed. Thirdly, since text may be surrounded by non-text CCs (e.g., the signboard containing characters is detected in Figure 1 (a)), we reject this kind of false positive by limiting the number of bounding boxes within a particular bounding box to three. For definitions of aspect ratio, saturation and area, see [12].

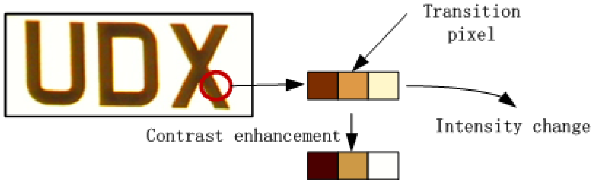
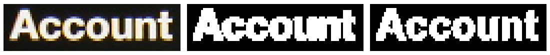


Figure 2. Contrast enhancement process.



(a) (b) (c)

Figure 3. (a) Original characters. (b) Bare MSER detection. (c) Contrast-enhanced MSER detection.

2.3. Stroke Width Extraction

Stroke width is defined as the length of a straight line from a text edge pixel to another along its gradient direction. The basic motivation of our stroke width extraction algorithm is that stroke width almost remains the same in a single character, however, there is significant change in stroke width in non-text regions as a result of their irregularity. There are several researches exploited this property, such as the work [1, 10], both of which calculate stroke width from a stroke boundary to another along gradient direction. Since skeleton is an effective tool to represent the structure of a region, inspired by the work [8] which uses skeleton to analyze text string straightness, we take advantage of skeleton to extract stroke width.

The initial step of stroke width extraction is to get skeletons of MSERs remained. On every foreground pixel on the skeleton, distance transform is applied to compute the Euclidean distance from this pixel to the nearest boundary of the corresponding MSER. Then we obtain a skeleton-distance map. This process is depicted in Figure 4. Figure 4 (a) illustrates a non-text MSER and text MSER from Figure 1 (a), and their corresponding skeleton and skeleton-distance map are shown in Figure 4 (b) and Figure 4 (c) respectively.

Variance on skeleton-distance map of each CC is

Table 1. Variances of false positive and characters.

False positive	'A'	'I'	'R'	'C'	'R'	'A'	'F'	'T'
0.4188	0.0978	0	0.0978	0.0594	0.0648	0.0833	0.0933	0.0706

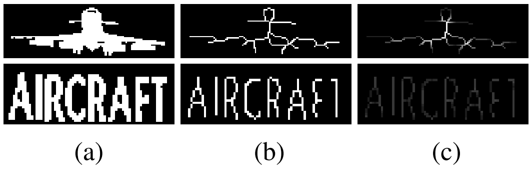


Figure 4. (a) Detected MSER of false positive and text. (b) Skeleton map. (c) Skeleton-distance map.

computed to measure the difference between text regions and false positives. Table 1 lists values of variance obtained from Figure 4 (c). Note that text characters have much smaller variances compared with the false positive. Based on this property we remove CCs with large variances. It can be seen in Figure 1 (c) that some false positives are eliminated after this procedure.

2.4. CC grouping

The main aim of CC grouping is to group adjacent characters detected in the previous steps into separated meaningful words and further reject false positives. Based on the observation that characters in the the same word usually share some similar properties, such as intensity, size, stroke width etc., these valuable information can be utilized in CC grouping. The details of our CC grouping method are illustrated below.

Center points of CCs are extracted as the first step of the proposed method. Then we obtain two maps, namely distance map and orientation map, by computing the Euclidean distance D and orientation angle θ between each CC pairs. If D is smaller than $MaxDistance$, which is defined as the maximum Euclidean distance from each CC to another, these two CCs are considered as adjacent candidates.

In the following step, we check θ between each adjacent pair of CCs on the orientation map. By assuming that texts usually lie in the horizonal direction, we set θ between $-30°$ and $30°$. Every pair of CC satisfying this rule is checked by similarity criteria below:

- w_i+w_j >1.2 × D
- max(w_i/w_j,w_j/w_i) <5
- max(h_i/h_j,h_j/h_i) <2
- max(s_i/s_j,s_j/s_i) <1.6
- max(n_i/n_j,n_j/n_i) <1.7

where w_i, h_i, s_i, n_i denote width, height, mean of stroke width, intensity of bounding box respectively, and all the thresholds are obtained from ICDAR 2003 training set. This is based on the observation that adjacent characters in the same word usually share similar stroke width and intensity. Adjacent CCs obeying all the rules are considered as true adjacent text characters thus are grouped together. The result of our CC grouping method is illustrated in Figure 1 (d), it is obvious that all characters are grouped successfully, meanwhile, all false positives are rejected.

3. Experiments

To evaluate the robustness of the proposed algorithm, we adopt the testing images in the public bench ICDAR 2003 text locating dataset [3] in our experiment. Three widely used measurement criterions, namely precision(p), recall(r) and f measure ($f = 1/(\alpha/p + (1 - \alpha)/r)$) [3] are exploited to evaluate the performance of our method. In order to detect both bright and dark text objects, two rounds of MSER detection are performed for each testing image and the final result is the combination of two round results.

As for the parameters setting, we set the gradient threshold T_1 as 30 and T_2 as 50 empirically. Besides, CCs whose stroke variance larger than 0.2 should be rejected. Furthermore, $MaxDistance$ is set as 300 to measure the maximum distance between two letters.

We compare our text detection result with a number of state-of-the-art methods tested on the same database using p, r and f criteria. The comparison result is shown in Table 2. We can see that the proposed approach has the highest recall rate of 0.59.

Recently, ICDAR 2011 Robust Reading Competition [7] was organized to evaluate the state-of-the-art process in text detection from complex nature scene. We also adopt the dataset used in this competition. Table 3 shows our text detection results on this dataset.

Figure 5 illustrates some results of our robust text detection algorithm. Estimated text regions are surrounded by blue bounding boxes. Note that the proposed method is insensitive to text color, font, size and position. With the proposed method, most text regions are detected, meanwhile, few false positives left.

We also present some failure examples in Figure 6. Because of the illumination problem, 'Bus' and 'Times' in Figure 6 (a) are not detected. All letters are discarded in Figure 6 (b) due to similar color between text and background. Moreover, characters 'X', 'M', and 'L' in Figure 6 (c) are eliminated because of large changes in

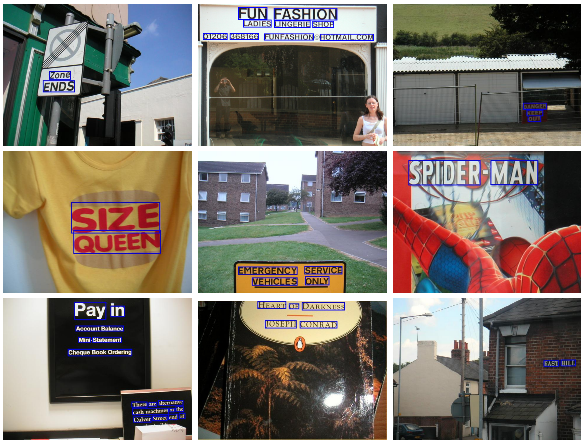


Figure 5. Sample output of our method.

Table 2. Result on ICDAR 2003 dataset.

Method	p	r	$f(\alpha = 0.5)$
Proposed	**0.59**	**0.59**	**0.59**
Neumann [5]	0.59	0.55	0.57
Zhang [11]	0.67	0.46	0.55
Liu [2]	0.66	0.46	0.54
Zhou[13]	0.57	0.50	0.53
Ashida [3]	0.55	0.46	0.50

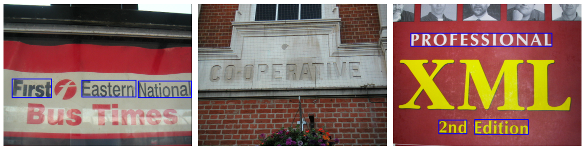
stroke width, but this kind of text is rare in the dataset, which will not affect the overall result to a large extent. We notice that the performance of our algorithm depends much on the potential text regions detected in the initial step (e.g., sometimes text cannot be detected using the contrast-enhanced MSER algorithm).

4. Conclusion

In this work, a novel CC-based methodology for text detection in natural scene images is presented. MSER-s are first utilized as potential text regions. A significant novelty of our work compared with previous research is that we apply skeleton to extract stroke width. Moreover, our robust CC grouping method can not only group characters into separated words, but also eliminate false positives at the same time. Text detection results on the ICDAR datasets demonstrate that our algorithm performs comparable to other methods.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (NSFC), No.61071209.

References



(a) (b) (c)

Figure 6. Failure examples.

Table 3. Result on ICDAR 2011 dataset.

Method	p	r	$f(\alpha = 0.5)$
Proposed	**0.59**	**0.62**	**0.61**
Neumann	0.69	0.53	0.60
TDM_IACS	0.64	0.54	0.58
LIP6-Retin	0.63	0.50	0.56
KAIST AIPR System	0.60	0.46	0.51
ECNU-CCG Method	0.35	0.38	0.37
Text Hunter	0.50	0.26	0.34

[1] B. Epshtein, E. Ofek, and Y. Wexler. Detecting text in natural scenes with stroke width transform. In *CVPR*, pages 2963–2970, 2010.

[2] Z. Liu and S. Sarkar. Robust outdoor text detection using text intensity and shape features. In *ICPR*, pages 1491–1496, 2008.

[3] S. M. Lucas, A. Panaretos, L. Sosa, A. Tang, S. Wong, and R. Young. Icdar 2003 robust reading competitions. In *ICDAR*, pages 682–687, 2003.

[4] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In *BMVC*, pages 384–393, 2002.

[5] L. Neumann and J. Matas. A method for text localization and recognition in real-world images. In *ACCV*, pages 770–783, 2010.

[6] Y. Pan, C. Liu, and X. Hou. Fast scene text localization by learning-based filtering and verification. In *ICIP*, pages 2269–2272, 2010.

[7] A. Shahab, F. Shafait, and A. Dengel. Icdar 2011 robust reading competition challenge 2: Reading text in scene images. In *ICDAR*, pages 1491–1496, 2011.

[8] P. Shivakumara, T. Q. Phan, and C. L. Tan. A laplacian approach to multi-oriented text detection in video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(2):412–419, 2011.

[9] X. Wang, L. Huang, and C. Liu. A new block partitioned text feature for text verification. In *ICDAR*, pages 366–370, 2009.

[10] J. Zhang and R. Kasturi. Character energy and link energy-based text extraction in scene images. In *ACCV*, pages 308–320, 2010.

[11] J. Zhang and R. Kasturi. Text detection using edge gradient and graph spectrum. In *ICPR*, pages 3979–3982, 2010.

[12] X. Zhao, K.-H. Lin, Y. Fu, Y. Hu, Y. Liu, and T. S. Huang. Text from corners: A novel approach to detect text and caption in videos. *IEEE Transactions on Image Processing*, 20(3):790–799, 2011.

[13] G. Zhou, Y. Liu, Z. Tian, and Y. Su. A new hybrid method to detect text in natural scene. In *ICIP*, pages 2653–2656, 2011.