An Improved Text Localization Method for Natural Scene Images

JIANG Mengdi1,* , CHENG Jianghua1 , CHEN Minghui2 and KU Xishu1
1 College of Electronic Science and Engineering, National University of Defense Technology, Changsha 410073, China
2 Military Representative Office of the PLA Rocket Force in Changsha Area, Changsha 410073, China
* jiangmengdi11@163.com

Abstract. In order to extract text information effectively from natural scene image with complex background, multi-orientation perspective and multilingual languages, we present a new method based on the improved Stroke Feature Transform (SWT). Firstly, The Maximally Stable Extremal Region (MSER) method is used to detect text candidate regions. Secondly, the SWT algorithm is used in the candidate regions, which can improve the edge detection compared with tradition SWT method. Finally, the Frequency-tuned (FT) visual saliency is introduced to remove non-text candidate regions. The experiment results show that, the method can achieve good robustness for complex background with multi-orientation perspective, various characters and font sizes.

1. Introduction

Text detection and localization in natural images serves is considered to be an important aspect of overall image understanding. It has gained considerable attention in both academia and industry in the last decade. Despite the tremendous effort devoted to solving this problem, text localization remains to be challenging due to unconstrained color, sizes, alignment of characters, lighting, various shapes of fonts and the complexity of scenes in natural images. Furthermore, text-like background objects, such as bricks, windows and leaves, often lead to many false alarms in text detection.

Existing text localization can be broadly classified into four main groups: edge based[1,2], texture based[3,4], connected component (CC)-based[5,6] and stoke based methods[7,8,9,10]. Among them, stroke based method has been shown to be effective and is heavily used in many recent approaches due to its insensitivity to the sizes, orientations, color, fonts and languages of the text. The stroke-based method utilizes a feature that distinguishes between the text and other elements of the scene - the nearly constant width of the stroke, which can be used to detect the area that contains text areas. The text can be regarded as a model of the combination of stroke elements in each direction, and the characteristics of the text can be extracted by the combination and distribution of stroke elements.

Jung C et al. [7] proposed the concept of Stroke filter (SF), which defines the characteristics of strokes. The article thinks that the features like edge (gradient), connected component, texture, and so on are the external features of the text, and the stroke is the intrinsic characteristic of the text, so the stroke-based method is adaptive not only to all languages but also to the handwritten text. Motivated by it, Epshtein et al.[8] proposed Stroke Width Transform (SWT). The algorithm first performs edge detection by canny operator, and searches for the opposite pixel in the edge gradient direction.
define the distance between the pair pixels as the stroke width, and use the feature of text strokes to
detect the text. This algorithm is dependent on the effect of edge detection. However, it’s hard to
obtain effective edge as the background is usually complex in natural scene images. Pan et al.[9]
presented a hybrid method for detecting and localizing texts in natural scene images by stroke
segmentation, verification and grouping. The author proposed a scale adaptive segmentation method to
extract stroke candidate and a CRF model with pair-wise weight by local fitting to verify stroke. Since
the stroke width transform relies on a precise edge detection scheme, Mosleh et al. [10] introduced a
novel bandlet-based edge detector which is quite effective at obtaining text edges in images as well as
dismissing noisy and foliage edges.

The SWT algorithm can detect most of the text area, but when the image background is complex, it
will produce serious false alarm, such as leaves, stripes, bushes, signs, houses which will often be
mistaken for candidate text area. The performance of the SWT algorithm highly depends on the edge
detector, which is often not feasible in many challenge cases. Compared to the SWT based methods,
MSER operator is capable of detecting more true text components. Therefore, we perform SWT
detection in the candidate regions detected by MSER in this paper. The maximally stable extremal
Regions contain most of the text and filter a lot of noise in the same time. So, perform SWT in the
regions detected by MSER can effectively improve the edge detection which is suitable for low
resolution text detection and localization. Moreover, the Frequency-tuned (FT) visual saliency is
introduced to improve the robustness of the algorithm to complex background. An important feature of
visual saliency is to achieve the suppression of background and the highlighting the significant areas.
With this property, we can remove the non-text regions effectively.

2. Related Technology.

2.1. Maximally Stable Extremal Region

The concept of Maximally Stable Extremal Regions (MSERs) was proposed by Matas et al. [11].
MSERs define an extremal region as a connected component of an image whose pixels have intensity
contrast against its boundary pixels. For a grayscale image, consider all possible thresholds, set the
pixels below the threshold to black, and set the pixels above the threshold to white. Set the threshold
small to large: $t = \{0, 1, \ldots, 255\}$, and obtain the corresponding images. First, we can see a full white
image. Then the image will slowly appear with some small grayscale black spots, and gradually
merged into small areas. In the last, the image will become full black. In the process of the increasing
of the threshold, the number of pixels within some areas is almost constant during the change of
several thresholds, this area is called the maximum stable extremal region (MSER).

Since the individual text in the scene image generally has equal color intensity and obvious
intensity contrast with the background. The MSER algorithm can effectively detect the text area in the
scene image. The test results are shown in figure 1:

![Figure 1. MSER method. (a) original image. (b) Maximally Stable Extremal regions. (c) candidate regions detected by MSER method](image)

2.2. Stroke Width Transform
The stroke width transformation is the method proposed by Boris Epshtein\[7\] in the 2010 CVPR conference. The text elements in the natural scene have an almost constant stroke width, and the stroke width of the text in its adjacent area is approximately equal, so we can use the stroke width of each pixel to obtain the text areas. The input of SWT is the original image, and the output is a stroke width image with a same size of original image, where the value of each pixel is its stroke width value. The process of SWT method is as follows: Firstly, initialize the value of SWT image to $\infty$. Then, use canny operator to detect the edge of the image, and each edge pixel has a gradient value in the gradient direction. Take an edge pixel $p$ (the gradient direction is $d_p$), and look for another edge pixel $q$ of the opposite gradient direction ($d_q$) of $d_p$. When $d_p$ and $d_q$ satisfied the condition: $|d_p + d_q| < \pi/6$, calculate the Euclidean distance $\|p - q\|$ between $p$ and $q$, and set all pixels on the path of $p$ and $q$ to this distance, that is, stroke width value. Figure 2 is the process of SWT method.

Figure 2. SWT: (a) A typical stroke. (b) Find the edge pixels $p$ and $q$. (c) Determine the stroke width.

If the width ratio of two adjacent pixels is within $[1/3, 3]$, the adjacent pixels are considered to be connected. Then, we filter the candidate regions follow the heuristic rules, and grouping letters into text lines.

2.3. Frequency-tuned (FT) visual saliency

Visual saliency analysis is a kind of innate ability of human beings, which is the basic function of visual information processing. Texts in the natural scenes images are mostly slogans, advertising and signs, and their visual perception is generally prominent and prominent, which is more attractive to attract people's attention. The Frequency-Tuned Saliency model is a kind of bottom-up saliency detection method proposed by Achanta et al. \[12\], and the saliency value of the pixel is obtained by a multi-scale method of the contrast of local color and luminance feature. The algorithm has three advantages over the existing method: 1) the salient regions have a clear boundary; 2) the saliency graph has full resolution; 3) the computational efficiency is high. The model is calculated in the Lab color space, which is defined as:

$$S(x, y) = \|I_{\mu} - I_{\text{adc}}(x, y)\|$$  \hspace{1cm} (1)

where $I_{\mu}$ is the mean image feature vector, $I_{\text{adc}}$ is the corresponding image pixel vector value after the Gaussian blur in Lab space (using a $5 \times 5$ separable binomial kernel) of the original image, and $\|$ is the $L_2$ norm. The FT visual saliency results are shown in figure 3.

Figure 3. FT visual saliency example: (a) original image (b) FT saliency image (c) saliency region segmented by FT saliency thresholding.
3. The proposed method
In this section, we describe the text detection algorithm. The key point of SWT is to precisely calculate the stroke width value of each pixel in the image, which depends on the good edge detection. In this paper, we first use MSER to detect candidate text areas, and perform SWT in the candidate areas, which will improve the effect of edge detection. Then, we use some heuristic rules to delete non-text areas. Finally, the FT visual saliency is introduced to further delete the false alarm areas. The algorithm has achieved good results, and its structure is shown in figure 4.

3.1. Preprocessing and MSER detection
The principle of the MSER algorithm has been introduced in 2.1. As the individual character in the scene image generally has uniform color intensity and obvious intensity contrast against background, the MSER method can effectively detect text regions. However, there still exist some problems: (1) For the low contrast image, the image is usually very vague, which makes difficult to extract the text region as the maximum stable extremal region. (2) For the image with non-uniform illumination, the effect of MSER extraction is poor. In this paper, histogram equalization is used to improve image contrast and contrast compensation is used to correct the effect of non-uniform illumination.

(1) Improve the low contrast image with histogram equalization. Histogram equalization is computationally fast, and the improved is shown in figure 5. The experimental results show that the MSER algorithm can detect the text area well in low contrast image after enhanced by histogram equalization.
Figure 5. From left to right: (a) original image and its MSER image and the text region detected by MSER. (b) the image enhanced by histogram equalization and its MSER image and the text region detected by MSER.

(2) Correct the influence of non-uniform illumination by contrast compensation. The algorithm first estimates the background image of the original image. We can simply understand that the uneven illumination background can be improved by the original figure minus the background map. However, it is conceivable that the contrast between the background and text is smaller when the background color is darker. Therefore, we will compensate the contrast according to the depth of the background color. As shown in figure 6, we can see that the text regions in original image can't be detected by MSER algorithm because of the non-uniform illumination. However, after the contrast compensation, the text region has been detected, and the false alarm areas will be deleted in the last step.

Figure 6. From left to right: (a) original image and its MSER image and the text region detected by MSER. (b) the image improved by contrast compensation and its MSER image and the text region detected by MSER.

3.2. Perform SWT in the candidate regions
The proposed method takes SWT in the candidate text regions detected by MSER algorithm. Since the MSER algorithm can locate most of the text areas and filter out a large number of background, the edge detection in MSER is significantly improved compared to the edge detection in the whole image. This step improves the accuracy of the calculation of the stroke width. Then we use SWT method to locate text regions and a series of heuristic rules to delete false alarm areas. The heuristic rules are as follows:

(1) The area of candidate regions. If the area of the candidate region is too small, the region is judged to be an alarm region. We set the region whose area is less than 20 pixels to be deleted.

(2) The aspect ratio of candidate regions. Many natural processes may generate long and narrow components that may be mistaken for possible letters. Therefore, we limit the aspect ratio to a value between 0.1 and 10 and the regions where the aspect ratio is too large or too small will be removed.
(3) The stroke width information of candidate regions. If the stroke width within a region changes too large, the regions should be removed. The text regions should satisfy the following conditions:

$$Var_{SWT} < \frac{Mean_{SWT}}{2}$$  \hspace{1cm} (2)

where $Var_{SWT}$ is the variance of stroke width in the candidate region; $Mean_{SWT}$ is the average stroke width in the candidate region.

The experimental results are shown in figure 7. We can see that the edge detection in the regions detected by MSER has been improved obviously. Compared with directly conducting SWT on the original images, the improved SWT location result is better with removing a non-text region. However, there are still some false alarm areas, and we will introduce FT visual saliency for the further judgments.

![Figure 7](image)

**Figure 7.** Comparison of SWT in whole image and candidate regions detect by MSER. From left to right: (a) original image and its edge image and the SWT location result. (b) MSER image and its edge image and the SWT location result.

### 3.3. Text location based on SWT and FT visual saliency

In this section, we proposed a method combined with SWT and FT visual saliency to further remove the false alarm regions. The texts in natural scene image mostly are slogans, advertising and signs which usually capture our attention. Thus the visual saliency value of text region is relative higher than the background. Therefore, we use this property to judge the candidate regions.

The value of each pixel in the visual saliency image indicates the saliency of the corresponding pixel of the original image, and the value is between 0 and 1. So we can set threshold to achieve the salient area detection. In short, if we get the visual saliency image, the pixels of salient object can be calculated by:

$$O(x) = \begin{cases} 
1 & \text{if } S(x) > T_s \\
0 & \text{otherwise} 
\end{cases}$$  \hspace{1cm} (3)

where $O(x)$ is the binarized image of the visual salient object regions; The threshold $T_s$ is based on empirical statistics, which is generally set as:

$$T_s = n \times E(S(x))$$  \hspace{1cm} (4)
Where \( n \) is the coefficient of \( T_s \) and the average saliency value. The principle based on SWT and FT visual saliency is described as follows: First, calculate the average saliency value of each candidate region \( SWT_i \) detected by SWT algorithm.

\[
\text{aveSaliency}_i = \frac{\sum S(x)}{\text{num}(x)}, \quad \forall x \in SWT_i(\bullet)
\]

(5)

where \( x \) is the pixel in the candidate region \( SWT_i \); \( S(x) \) is the FT visual saliency image. If the average saliency value of the candidate region is greater than the threshold, it is retained. Otherwise, the candidate region whose average saliency value is less than the threshold will be removed.

\[
SWT_i = \begin{cases} 
\text{text regions} & \text{if } \text{aveSaliency}_i \geq T_s \\
\text{false alarm regions} & \text{otherwise}
\end{cases}
\]

(6)

4. Experiments and results

The performance of the proposed method was evaluated on the ICDAR'11/13 dataset, which remains the most widely used benchmark for text detection in natural scenes. The performance of our method is quantitatively measured by precision (P), recall (R) and F-measure (F), which are defined as follows:

\[
R = \frac{C}{T}
\]

(7)

\[
P = \frac{C}{E}
\]

(8)

\[
F = \frac{1}{\frac{\alpha}{P} + \frac{1-\alpha}{R}}
\]

(9)

where \( C \) is the number of the correctly detected text regions, \( T \) is the number of the actual text regions, and \( E \) is the number of the detected text regions (including the false alarm regions). F-measure is the weighted average of precision and recall, where \( \alpha \) is the weighting factor, and in this paper. The algorithm is compared with the classic SWT and MSER. The results are shown in Table 1. The evaluation of the first two algorithms is measured on the ICDAR'03/05 dataset. The algorithm is measured in the ICDAR'11/13 dataset. The ICDAR'11/13 dataset is extended on the ICDAR'03/05 dataset, containing multi-directional and distorted scene text images, which is more difficult to detect the text. From Table 1 we can see that this method has high extraction accuracy and good overall performance. Figure 8 shows some comparison of the results of classic SWT algorithm the proposed method.

| author     | P   | R   | F   | property               |
|------------|-----|-----|-----|------------------------|
| Epshtein[9]| 0.73| 0.60| 0.66| SWT                    |
| Matas[13]  | 0.59| 0.55| 0.57| MSER                   |
| Our method | 0.78| 0.66| 0.71| MSER, SWT, FT visual saliency |
5. Conclusion
In this work we have presented a novel system for detecting and localizing text-lines in natural scene images. First, we preprocess the image to improve the influence of low contrast and non-uniform illumination, which makes the subsequent MSER locate more accurate. Second, we conduct SWT in the candidate text regions detected by MSER. Finally, we introduce FT visual saliency to remove non-text candidate regions. Experimental results show that our approach has achieved the state-of-the-art results on ICDAR’11/13 dataset.

References
[1] Q. Ye, J. Jiao, J. Huang, H. Yu. Text detection and restoration in natural scene images [J]. Journal of Visual Communication and Image Representation, 2007, 18(6) 504-513.
[2] X. Liu, J. Samarabandu. Multiscale edge-based text extraction from complex images [C]//International Conference on Multimedia and Expo, IEEE, 2006:1721–1724.
[3] Y. F. Pan, C. L. Liu, X. Hou. Fast scene text localization by learning-based filtering and verification [C]// 17th IEEE International Conference on Image Processing ,ICIP, IEEE, 2010:2269-2272.
[4] G. Zhou, Y. Liu, Q. Meng, Y. Zhang. Detecting multilingual text in natural scene [C]// International Symposium on Access Spaces, ISAS, IEEE, 2011:116–120.
[5] H. Zhang, C. Liu, C. Yang, X. Ding, et al. An improved scene text extraction method using conditional random field and optical character recognition [C]//International Conference on Document Analysis and Recognition ,ICDAR, IEEE, 2011:708-712.
[6] Y. Pan, X. Hou, C. Liu. A hybrid approach to detect and localize texts in natural scene images [J]. IEEE Transactions on Image Processing, 2011,20:800-813.
[7] Jung C, Liu Q, Kim J.A stroke filter and its application for text localization [J]. Pattern Recognition Letters, 2009,30(2):114-122.
[8] Epshtein B, Ofek E, Wexler Y. Detecting text in nature scenes with Stroke Width Transform [C]//Proceedings of Computer Vision and Pattern Recognition, CVPR, IEEE,2010:2963-2970.
[9] Pan Y F, Zhu Y, Sun J, et al. Improving Scene Text Detection by Scale-Adaptive Segmentation and Weighted CRF Verification[C]// International Conference on Document Analysis and Recognition. IEEE, 2011:759-763.
[10] Mosleh A, Bouguila N, Hamza A B. Image text detection using a bandlet-based edge detector and stroke width transform [C]// British Machine Vision Conference, 2012.
[11] L. Neumann, J. Matas. A method for text localization and recognition in real world images [J]. Computer Vision, ACCV, 2011: 770-783.
[12] Achanta R, Hemami S, Estrada F, et al. Frequency-tuned salient region detection[J]. 2009, 22(9-10):1597-1604.