15th IAPR International Conference on Machine Vision Applications (MVA)
Nagoya University, Nagoya, Japan, May 8-12, 2017.

Scene Text Extraction with Local Symmetry Transform

Qi Chen
Xi'an Jiaotong University
Xi'an 710049, China
qzachenqi@stu.xjtu.edu.cn

Yonghong Song
Xi'an Jiaotong University
Xi'an 710049, China
songyh@mail.xjtu.edu.cn

Yuanlin Zhang
Xi'an Jiaotong University
Xi'an 710049, China

Abstract

As an important part of scene understanding, scene text extraction can promote the performance of text recognition significantly. This paper focuses on extracting text pixels from background in images of natural scenes. First, a novel feature descriptor called local symmetry transform is proposed to detect local symmetry relationship of strokes for each image pixel. The presented method is local and color invariant, which makes it robust enough to handle many complex situations for scene text extraction such as variation of lighting, blur of text regions, shadow and complex background. Then, guided filter is introduced to enhance the local symmetry map produced by the local symmetry transform, which can smooth the feature map and reduce noise noticeably. Finally, text candidates are generated as seed points, and by seed-based segmentation we automatically judge the text polarity to obtain correct text regions. Experimental results on KAIST dataset show that our method leads to a high performance.

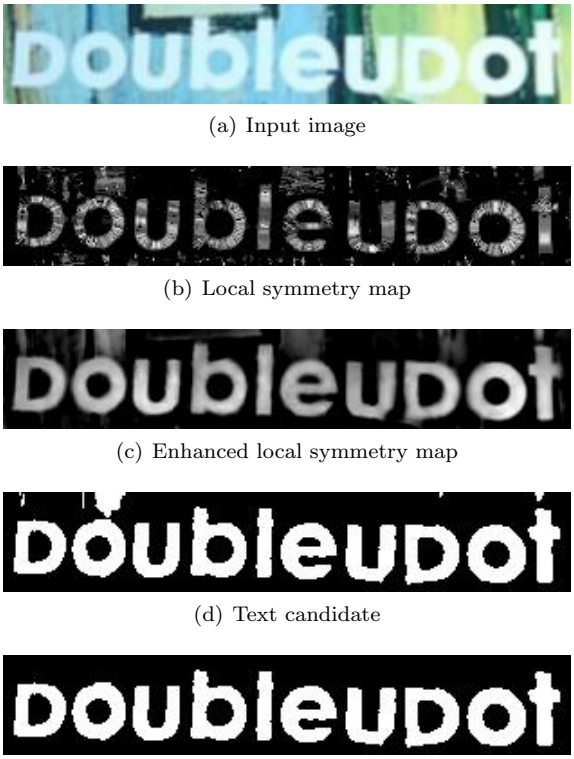
1 Introduction

As an important part of natural scene image, textual information in scene image can always offer core clues for people or computers to understand scenes. Hence, to obtain textual information in natural scene image is a valuable and promising research direction, which can be utilized in many ways.

In recent years, scene text extraction algorithms have been classified into three categories: adaptive threshold based, probability model based, and clustering based. Adaptive threshold based method, holding the assumption that color in the text area is similar, segments images according to local color features. For instance, OTSU [1] , Kittler [2], Niblack [3] and Sauvola [4]. The main drawback of adaptive threshold based method is that it is difficult to choose a effective threshold value to extract text, especially when text is blurred.

Probability model based method tends to establish a probability model for image, which transforms text extraction problem into a pixel labeling problem by minimizing a carefully designed energy function. Lee and Kim [5] proposed employing a two-stage CRF model to label groups of text regions.

Taking the advantage of the characteristic that text regions tend to possess similar color and texture, clustering based method segments text and background by clustering image pixels. Wakahara Toru and Kita Kohei [6] used K-means method to generate text candidates and then used support vector machines (SVM) to classify text candidates. After merging the results of classification, text regions were extracted.



(a) Input image

(b) Local symmetry map

(c) Enhanced local symmetry map

(d) Text candidate

(e) Result

Figure 1. The pipeline of our method. The input image is shown in (a), local symmetry map (b) is generated by applying LST on the input image (a), then we apply guided filter on local symmetry map (b) to obtain a enhanced local symmetry map (c), next text candidate (d) is generated by using OTSU threshold method on (c). After text polarity estimation and denoising, the segmented text is shown in (e).

The approach we proposed in the study is inspired by symmetry detection and stroke width transform (SWT) [7], which is proved to be effective, proposes detecting stroke pixels by measuring the orientation difference between pairs of edge pixels, and grouping stroke pixels with similar widths as connected components. Daniel Reisfeld [8] associated symmetry magnitude and direction with each pixel in an image to compute the symmetry map directly from the edge map of images without prior segmentation.

In this paper, we propose a novel text extraction method, Firstly local symmetry transform (LST) is used to generate local symmetry map and stroke width map simultaneously. Then we take the advantage of weighted median filter [9] and guided filter [10] to enhance stroke width map and local symmetry map respectively. After that, we utilize adaptive threshold

method OTSU to generate text candidates. Finally image segmentation method random walk [11] is used to automatically detect text polarity and to choose the correct text region from text candidates. The pipeline of our approach is shown in Fig. 1. It makes the following contributions:

1. Combining generalized symmetry transform with widely-used SWT, we present LST to detect bilateral symmetry of the symmetric edge of strokes, and stroke width map is generalized simultaneously;

2. Weighted median filter and guided filter are introduced into scene text extraction. The experimental results of our approach have shown that these image filters can improve the feature maps commendably.

2 proposed method

Text extraction is usually employed after text localization. The input of our text extraction module is the localized color image of text region. In order to improve the final performance, we use pre-process to improve image quality and to reduce calculating time consumption.

Firstly, we use Gaussian filter to smooth input image and then the input image is normalized to a fixed scale, namely scaling the image so that the minimum of image height and width is a fixed value. Afterwards LST is used to generate local symmetry map and stroke width map. Usually the result of LST is not good enough due to noise and complex background of scene image. Therefor image filtering is applied to enhance the output of LST, where we use weighed median filter [9] to refine stroke width map and guided filter [10] to refine local symmetry map. Next, OTSU method is employed on local symmetry maps to generate text candidates. Because there are two polarities of point, namely along the gradient direction and against the gradient direction, two sets of text candidates are obtained. Finally, an image segmentation based method is used to automatically detect text polarity and to choose the correct text regions from the two sets of text candidates.

2.1 Local Symmetry Transform

Notice that the strokes of text usually have bilateral symmetric relationship[12]. Here, we present LST to detect local bilateral symmetry of text strokes. LST is designed so that text strokes with more obvious bilateral symmetric relationship have a higher value, while the value of background region is relatively low. The input of LST is the localized color image of text region after pre-processing. Firstly, we convert the input image into a gradient map using Scharr operator [13], then the gradient map is normalized into $[0,1]$,and the gradient direction is normalized into $[-\pi,\pi]$.

Then we define a local symmetry measure for each point pair, let $p_k = (x_k, y_k)$ be any point $(k = 1, 2, \ldots, K)$, K means the maximum index of point on the gradient map, $g(p_k)$ and $\theta(p_k)$ denote the gradient strength and gradient direction of point p_k.

The local symmetry measure function is defined as:

$$R(p_i,p_j)_{\pm} = \begin{cases} O(p_i,p_j) * C(p_i,p_j) * G(p_i,p_j) * D(p_i,p_j), O(p_i,p_j) > ot \\ 0, others \end{cases} \quad (1)$$

Here, $p_j = p_i \pm w * (\cos(\theta(p_i)), \sin(\theta(p_i))) \; s.t. \; w < wt$. p_j is the next point on the gradient direction of p_i, wt is a stroke width threshold denotes the maximum stroke width our method can deal with, ot is a threshold to remove some noises and reduce calculation time consumption, the first function O measures the direction symmetry of the two points, it is defined as:

$$O(p_i,p_j) = P(\theta(p_i), \theta(p_j)) * max(P(\theta(p_i), \theta_l), P(\theta(p_j), \theta_l)) \quad (2)$$

Where P denotes the direction opposition of two angles, and θ_l means the direction of the line segment l passing through the two point p_i and p_j, the function P is defined as:

$$P(\theta_1, \theta_2) = \max(0, \cos(abs(\theta_1 - \theta_2) - \pi)) \quad (3)$$

Then C function measure the color consistency between the two points, it will be relatively high if the colors of points on the line segment l are almost all the same. The item is defined as:

$$C(p_i,p_j) = \max\left(0, \frac{(\min(g(p_i),g(p_j)) - \max(g(p_k)))}{\min(g(p_i),g(p_j))}\right), p_k \in L \quad (4)$$

Here, L is the points set of the line segment l without end points. Next the G function measures the gradient similarity of the two points, it is defined as:

$$G(p_i,p_j) = \max\left(0, 1 - \frac{abs(g(p_i) - g(p_j))}{\min(g(p_i), g(p_j))}\right) \quad (5)$$

Finally, the last function D is a distance weight function measures the distance proximity, and it is defined as:

$$D(p_i,p_j) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\|p_i - p_j\|}{2\sigma}\right) \quad (6)$$

Here, σ with different values implies different scales, and it is chosen based on experience.

While traversing on the image, for each point pair (p_i, p_j), all the points belonged to L are assigned to the local symmetry value $R(p_i, p_j)$. For each point, it can be inferred that the point may be assigned to several different local symmetry values, we only keep the maximum. And then the stroke width is defined as the length of line segment l, the same as before, for each point, we only keep the stroke width while the local symmetry value of the point is the maximum.

As mentioned before, there are two polarities of point, namely along the gradient direction and against the gradient direction. We use subscript $\pm$ to indicate different polarities in (1). We assume that the polarity of text stroke is consistent in an image, it is acceptable in most instances.

Taking advantage of the property of polarity, we use a strategy called polarity constrain to segment the image initially. Let $S(p_k)_{\pm}$ define the positive/negative local symmetry value of the point p_k, and $\widehat{S(p_k)}_{\pm}$ denotes the corresponding local symmetry value after applying polarity constrain, we modify the local symmetry value as follow:

(a) Input image (b) Positive local symmetry map (c) Negative local symmetry map

Figure 2. An example of LST. Input image is shown in (a). After using LST and polarity constrain, a positive local symmetry map (b) and a negative local symmetry map (c) are generated.

$$\widehat{S(p_k)}_+ = \max(0, S(p_k)_+ - S(p_k)_-) \quad (7)$$

$$\widehat{S(p_k)}_- = \max(0, S(p_k)_- - S(p_k)_+) \quad (8)$$

Finally, we get a positive local symmetry map and a negative local symmetry map, and the two local symmetry maps give the input image an initial segmentation as shown in Fig. 2.

2.2 Local Symmetry Map Enhancement

As shown in Fig. 2 (b)(c), the local symmetry map is not smooth because the local symmetry value of point on the image is assigned line by line, and there may exist isolated noise due to complex background. Besides, The same as SWT, in more complex situations, like corners, the LST values will not get correct symmetry value. SWT solves this problem by passing along each non-discarded ray again and assigning the median SWT value of all the pixels on the ray to all its pixels. We can not take example by SWT, because LST is applied on gradient map, there is no non-discarded ray defined in LST. We solve this problem by applying guided filter on local symmetry map.

First, we use weighted median filter [9] to refine the stroke width map. Based on original median filter, weighted median filter is first used for disparity refinement. For each pixel, every pixel in the neighborhood is weighted by using filter weights.

Then, we use guided filter [10] to smooth the local symmetry map, guided filter is an edge-preserving smoothing operator, and it generates the filtering output by considering the content of a guidance image. We use origin color image as the guided image, half of w_0 is set as the local window radius, and w_0 is the median of the corresponding stroke width map. After filtering, we use polarity constrain again.

2.3 Text Polarity Estimation

As instructed before, we have obtained two enhanced local symmetry map $S(p_k)_+$ and $S(p_k)_-$ after LST and enhancement, adaptive threshold image segmentation method OTSU [1] is used to generate text candidates on both enhanced local symmetry maps.

After that, we have to estimate the polarity of text regions, and figure out which one of the two sets of text candidates is foreground and which one is background.

Seed based image segmentation method is used to segment the whole image, the two sets of text candidates are considered as two different sets of seed points, we assign different labels to these seed sets respectively. Many seed-based segmentation methods can be used, we have considered the following methods: graph cut [14] , region grow [15] and random walk [11]. We choose random walk due to its good performance and high efficiency. Random work is performed in a 3-dimensional space by representing each pixel using pixel intensity, positive local symmetry value and negative local symmetry value. After segmentation, the whole image is segmented into two parts, then we calculate the number of points of each label on the image's border, the label which contains more points on the image's border is regarded as background, and we only retain the foreground.

2.4 Denoising

Because of the complex situation in scene text image, there are still some noises after we get the text regions. Here, we use some connected component features to filter out the noises. The connected components which satisfy the following constrain are regarded as noises:

$$CC\_minor\_axis < CC\_stroke\_width \quad (9)$$

$$CC\_stroke\_width < t_1 \quad (10)$$

$$CC\_top > t_2 \times image\_height \quad (11)$$

$$CC\_bottom < t_3 \times image\_height \quad (12)$$

Here, CC_minor_axis means the connected component's minor axis length which is calculated by ellipse fitting, CC_stroke_width is the mean value of the connected component's stroke width which can be calculated on the stroke width map, CC_top is the y coordinate of the upper-left corner of the connected component's bounding box, CC_bottom is the y coordinate of the lower-right corner of the bounding box, and $image\_height$ is the height of the image. t_1 , t_2 and t_3 are some empirical values.

3 Experimental results

The proposed text extraction method is evaluated by the English subset of Korea Advanced Institute of Science and Technology (KAIST) scene text database [16], which is widely used as experimental dataset for text extraction in natural images. Due to the de-nosing procedure, our method cannot solve multi-line text situation, so we select 493 single text line images from entire KAIST dataset.

For our experiments, we try to find a set of parameters applied to all images. For each image, we enlarge or shrink the image size proportionally so that the minimum of image height and width is fixed to 48 pixels. Stroke width threshold wt is set as 16 so that we can handle most of scene text images while direction symmetry threshold ot is set as 0.5 so as to remove some

Table 1. Experimental results on the KAIST dataset

Method	P	R	F
Our method	**0.92**	**0.86**	**0.89**
OTSU	0.72	0.94	0.82
Niblack	0.68	0.97	0.80
Sauvola	0.65	0.97	0.78
K-means	0.74	0.96	0.84
Bilateral Regression [a]	0.67	0.88	0.76
Bilateral Regression [b]	0.78	0.94	0.85

[a] original bilateral regression approach.
[b] bilateral regression using our text polarity estimation method.

false point pairs which don't seem to be point pairs of strokes. The window radius for weighed median filter is set as 2 empirically, and according to the suggestion in [10] we choose 0.0001 as the regularization parameter ε of guided filter. The parameters for denoising is chosen from experience t_1 , t_2 and t_3 are set as 2.5, 0.7 and 0.3 respectively.

The performance of our text extraction method is evaluated by *precision(P)*, *recall(R)* and *F-measure(F)*. They are measured using the same definitions in [17, 18] on pixel level. The performance of the proposed approach is shown in Table 1, The *precision*, *recall* and *f-measure* of our method are 0.92, 0.86 and 0.89 respectively.

The performance of the proposed method is compared with adaptive global threshold method OTSU [1], local threshold method niblack [3] and sauvola [4], clustering method k-means and bilateral regression method [19, 20]. All the results are shown in Table 1. However, the comparison methods can't estimate the text polarity by themselves except for bilateral regression method. Therefore, after binarization, we add our text polarity estimation procedure for OTSU, niblack, sauvola and k-means methods.

Our system is implemented in MATLAB, and the average execution time is 15 seconds. LST consumes most of the execution time.

4 Conclusions

We have proposed a novel method for text extraction in scene image. Combining generalized symmetry transform with widely-used SWT, we present LST to detect local bilateral symmetry of strokes. LST is firstly used to generate local symmetry map and stroke width map. Then we introduce image filter which utilizes the structures in the guidance image into scene text extraction to improve the feature maps. Our approach is evaluated by the KAIST scene text dataset and experimental results show that our approach has achieved outstanding performance.

References

[1] N. Otsu, "A threshold selection method from gray-level histograms," *Automatica*, vol. 11, no. 285-296, pp. 23–27, 1975.

[2] J. Kittler, J. Illingworth, and J. Föglein, "Threshold selection based on a simple image statistic," *Computer vision, graphics, and image processing*, vol. 30, no. 2, pp. 125–147, 1985.

[3] W. Niblack, *An introduction to digital image processing*. Strandberg Publishing Company, 1985.

[4] J. Sauvola and M. Pietikäinen, "Adaptive document image binarization," *Pattern recognition*, vol. 33, no. 2, pp. 225–236, 2000.

[5] S. Lee and J. H. Kim, "Integrating multiple character proposals for robust scene text extraction," *Image and Vision Computing*, vol. 31, no. 11, pp. 823–840, 2013.

[6] T. Wakahara and K. Kita, "Binarization of color character strings in scene images using k-means clustering and support vector machines," in *Document Analysis and Recognition (ICDAR), 2011 International Conference on*, pp. 274–278, IEEE, 2011.

[7] B. Epshtein, E. Ofek, and Y. Wexler, "Detecting text in natural scenes with stroke width transform," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pp. 2963–2970, IEEE, 2010.

[8] D. Reisfeld, H. Wolfson, and Y. Yeshurun, "Context-free attentional operators: the generalized symmetry transform," *International Journal of Computer Vision*, vol. 14, no. 2, pp. 119–130, 1995.

[9] Z. Ma, K. He, Y. Wei, J. Sun, and E. Wu, "Constant time weighted median filtering for stereo matching and beyond," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 49–56, 2013.

[10] K. He, J. Sun, and X. Tang, "Guided image filtering," in *Computer Vision–ECCV 2010*, pp. 1–14, Springer, 2010.

[11] L. Grady, "Random walks for image segmentation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 28, no. 11, pp. 1768–1783, 2006.

[12] T. Q. Phan, P. Shivakumara, and C. L. Tan, "Text detection in natural scenes using gradient vector flow-guided symmetry," in *International Conference on Pattern Recognition*, pp. 3296–3299, 2012.

[13] H. Scharr, *Optimal operators in digital image processing*. PhD thesis, 2000.

[14] Y. Y. Boykov and M.-P. Jolly, "Interactive graph cuts for optimal boundary & region segmentation of objects in nd images," in *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, vol. 1, pp. 105–112, IEEE, 2001.

[15] R. C. Gonzalez and R. E. Woods, "Digital image processing," 2002.

[16] J. Jung, S. Lee, M. S. Cho, and J. H. Kim, "Touch tt: Scene text extractor using touchscreen interface," *ETRI Journal*, vol. 33, no. 1, pp. 78–88, 2011.

[17] B. Bai, F. Yin, and C. L. Liu, "A seed-based segmentation method for scene text extraction," in *Document Analysis Systems (DAS), 2014 11th IAPR International Workshop on*, pp. 262–266, IEEE, 2014.

[18] M. S. Cho, J.-H. Seok, S. Lee, and J. H. Kim, "Scene text extraction by superpixel crfs combining multiple character features," in *Document Analysis and Recognition (ICDAR), 2011 International Conference on*, pp. 1034–1038, IEEE, 2011.

[19] J. Feild and E. G. Learned-Miller, "Scene text recognition with bilateral regression," *UMass Amherst Technical Report*, 2012.

[20] J. L. Feild and E. G. Learned-Miller, "Improving open-vocabulary scene text recognition," in *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*, pp. 604–608, IEEE, 2013.