Text Localization in Natural Images Using Discriminative Local Color Information

Chien-Cheng Lee and Shang-Fei Shen
Department of Communications Engineering, Yuan Ze University, Taoyuan, Taiwan
Email: cclee@saturn.yzu.edu.tw

***Abstract*—In this paper, we propose a new method of using local color information for text localization in natural images. The FAST salient point detector was applied in the first step to detect salient points from natural images. Then, the local color information of each salient point was extracted. The salient points was transformed into a bi-modal space to discover the text property of these points. After that, a clustering algorithm was used to find clusters of text area on images and got the possible positions of the text area. The distribution density was estimated and the average distance of salient points was calculated on the possible text area. Finally, minimum bounding box of the text area was defined to represent the text region. Experimental results have shown the advantages and effectiveness of the proposed method in the text detection in natural images.**

***Index Terms*—feature point detection, text detection, natural images**

I. Introduction

Text recognitions in natural images can be used in many applications in smart life. These techniques help machines to sense the human society and understand surrounding things. Text detection and localization are important preprocessing steps before text recognition. Text detection roughly classifies text and non-text regions. Text localization determines the accurate boundaries of text strings. After the text areas are localized, the Optical Character Recognition (OCR) software can recognize the text easily.

Many methods have been proposed for text detection and localization. Most of them are texture-based [1, 2] , edge-based [3, 4], or connected component-based [5, 6] methods. However, many difficulties are still needed to be solved, such as complex background, unknown text color, and degraded text quality. These difficulties emerge significantly in natural scene images because there are many uncertainties in the natural scene.

In this study, we focus on natural scene images. We proposed a novel feature point based text detection approach to locate the text area for natural images. The idea of our proposed approach is that the color of text is usually different from the background color so the text can be recognized by human eyes easily. Based on this idea, several feature points on images are extracted, and then the color distributions of these points are analyzed. If the color distributions contain the characteristic of text, the corresponding feature points are grouped and formed a candidate block of text area. Finally, a following image processing procedures was used to obtain the complete minimum bounding box of the text area.

The remainder of the paper is organized as follows. In Section II, we presented our feature point based text detection method of natural images. Comparison results are presented in Section III. Finally, in Section IV, we draw conclusions.

II. Methods

Our approach can be subdivided into three major steps: 1) feature point detection, 2) feature point clustering, and 3) minimum bounding box localization. A well-known algorithm, FAST (Features from Accelerated Segment Test) salient point detector, was used in the first step.

A. Color Space Transform and Feature Point Detection

To facilitate the discrimination of the text from the background, the color space of an image was converted from RGB to HSV color space first. In HSV color space, hue can be used to differentiate different colors, but it cannot indicate the weak color. That is, if a pixel has similar values in R, G, and B, hue will approach zero.

In natural images, illumination changes can influence the color representation of images, and make the color looks like gray. In this paper, we re-defined the hue transformation in the RGB to HSV color transformation. The new hue transformation is defined as (3).

$$V = max(R, G, B), \tag{1}$$

$$S = \begin{cases} 0, if\ V = 0 \\ \frac{V - min(R,G,B)}{V}, \end{cases} \tag{2}$$

$$H_{new} = \begin{cases} -1 & , if\ V - min(R,G,B) < T_a \\ 60° * \left(\frac{G-B}{V-min(R,G,B)}\right) + 0° & , if\ R = V\ and\ G \geq B \\ 60° * \left(\frac{G-B}{V-min(R,G,B)}\right) + 360° & , if\ R = V\ and\ G < B \\ 60° * \left(\frac{B-R}{V-min(R,G,B)}\right) + 120° & , if\ G = V \\ 60° * \left(\frac{R-G}{V-min(R,G,B)}\right) + 240° & , if\ B = V \end{cases} \tag{3}$$

where T_a is a pre-defined threshold which indicates the pixel color is weak.

Manuscript received June 16, 2017; revised November 7, 2017.

doi: 10.18178/ijsps.5.3.89-93

After all image pixels are transformed according to (1)-(3), the hue values marked as -1 will be assigned as the mean value of its neighbors according to a post hue transform algorithm described in Fig. 1.

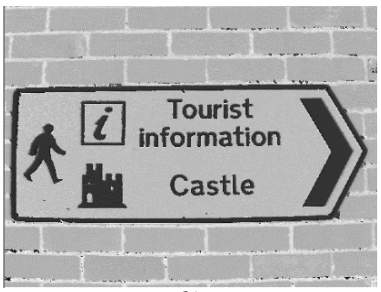
Algorithm - post hue transform algorithm

```
1:    Let H be the hue image
2:    Let B(i) be the neighbors of pixel i in H
3:    for each pixel i in H {
4:      if( H(i) == -1) {
5:        n = 0
6:        average = 0
7:        for each pixel k in B(i){
8:          if(k != -1) {
9:            average += H(k)
10:           n++
11:         }
12:       }
13:       H(i) = average/n
14:     }
15:   }
```

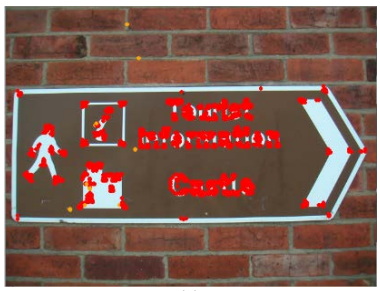
Figure 1. Post hue transform algorithm.

Then, a FAST salient point detector was applied in the R, G, B, hue, and value five channel images to extract feature points. These feature points are combined, and clustered in the next step. An example result of the color space transformation and feature point detection is shown in Fig. 2.

(a)



(b)



(c)

Figure 2. Example result of the color space transformation and feature point detection. (a) original image, (b) hue image of HSV color space, (c) feature points extracted from the hue image.

B. *Feature Point Clustering*

After the feature point extraction, we computed the color histogram for each feature point by selecting a $N \times N$ block centered around the point. Next, kernel density estimation with Gaussian kernel was used to smooth the histogram to find a smoother histogram. An example is shown in Fig. 3. Observing the figure, the maxima of the histogram could be identified easily. If the feature point belongs to text, the smoothed color histogram should have a bi-modal shape, as shown in Fig. 4. Otherwise, the feature points without bi-modal shape indicate these points do not belong to text area.

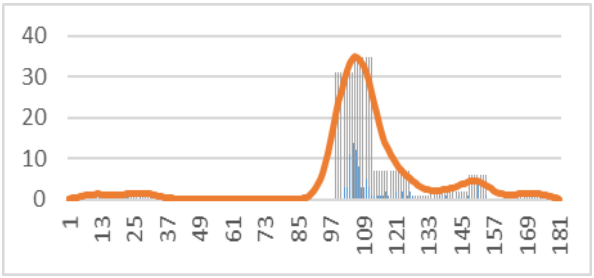


Figure 3. Example of color histogram. The smooth solid line is the result of kernel density estimation.

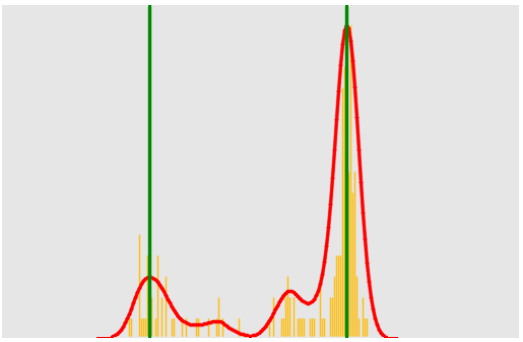
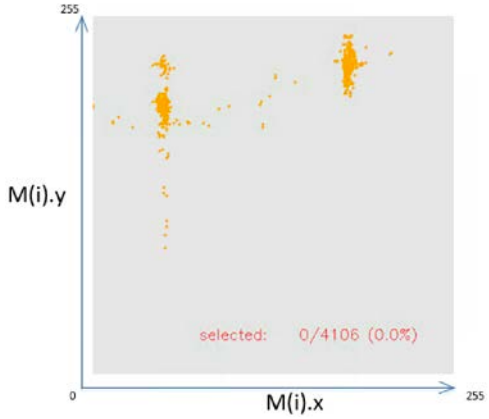


Figure 4. Bi-modal shape.

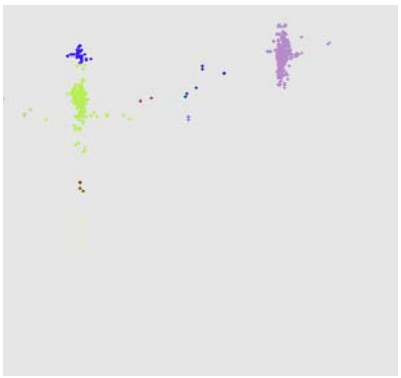
To cluster the feature points into text and non-text clusters, we transform the feature points from the image coordinate space to a bi-modal space. Let $k_1(i)$ and $k_2(i)$ be the histogram bin numbers of the highest and the second highest peaks of the smoothed histogram of feature point p_i. The point set $M(i)$ is defined as:

$$M(i) = \begin{cases} point\left(k_1(i), k_2(i)\right), if\ k_1(i) < k_2(i) \\ point\left(k_2(i), k_1(i)\right), if\ k_1(i) \geq k_2(i) \end{cases} \quad (4)$$

From the map of the point set $M(i)$, it can be observed that the feature points form a cluster when these points have the similar bi-modal histogram, as shown in Fig. 5(a). That is, these feature points have the similar foreground and background color in the image. In other words, these points belong to the same text block. In this study, gravitational clustering algorithm [7, 8] was used to cluster the point set. The clustering result is shown in Fig. 5(b). The feature points in the most compact cluster belong to the same text block. These points were superimposed on the original image and are shown in Fig. 5(c).



(a)



(b)

(c)

Figure 5. Feature points clustering. (a) Point set $M(i)$. (b) Clustering result. (c) Superimposed image of the original image with the most compact cluster points.

C. Minimum Bounding Box Localization

From Fig. 5(c), we can see that most of these points construct the text block. However, there are some unwanted points outside the text block. To remove the points outside the text block and find the minimum bounding box of the text block, the well-known segmentation method of vertical and horizontal projection was used in this study. The vertical and horizontal projection histograms were denoted by $V(x)$ and $H(y)$, and the value indicated the number of feature points in the Column(x) and Row(y), respectively, as shown in Fig. 6(a). Then, a text element T with width T_w and height T_h is defined as:

$$T_w = \frac{\Sigma_{x=R_{i1}}^{R_{i2}} V(x)}{R_{i2}-R_{i1}} \quad (5)$$

$$T_h = \frac{\Sigma_{y=C_{j1}}^{C_{j2}} H(y)}{C_{j2}-C_{j1}} \quad (6)$$

where interval $[R_{i1}, R_{i2}]$ represents the continuous row segment i with $V(x) > \delta$, interval $[R_{i1}, R_{i2}]$ represents the continuous column segment j with $H(y) > \delta$, and δ is a pre-defined threshold.

Using the text element, a text map image I is determined as follows:

$$I(x,y) = \begin{cases} 255, & pixel\ (x,y) contained\ in\ T\ centered\ around\ p(i) \\ 0, & otherwise \end{cases} \quad (7)$$

Next, a morphological closing operation with 5×5 structuring element was performed on the text map image five times, as shown in Fig. 6(b). Finally, minimum bounding boxes were localized for the connected components which their sizes were greater than a given threshold. The final result is shown in Fig. 6(c).

III. Experimental Results

The proposed method was evaluated on two datasets, the ICDAR 2013 and our collection from streets in Taiwan. The performance of our method is quantitatively measured by precision, recall, and Hmean. In the ICDAR dataset, our performance was medium. The recall, precision, and Hmean evaluation scores were about 0.7. However, we used the same evaluation to test the Taiwan street scene dataset. The scores were about 0.8. We also compared with the BUCT_YST algorithm. Table I and II show the comparison results of our method and the BUCT_YST method for these two datasets. Some results are shown in Fig. 7.

The comparison results show that our method demonstrates good performance in Taiwan street scene images, but gets medium performance in ICDAR dataset. It seems reasonable because most texts in ICDAR dataset were western characters. The western characters contain less junction, turning, and end points than Chinese characters. Hence, the number of feature points in western characters is less than that in Chinese characters. These results lead to medium performance. In contrast, BUCT_YST algorithm got low scores in Taiwan street scene images. It seems that the BUCT_YST algorithm is based on neural network training. The performance degradation came from the unseen in the training phase.

TABLE I. Performance Comparison for ICDAR Dataset

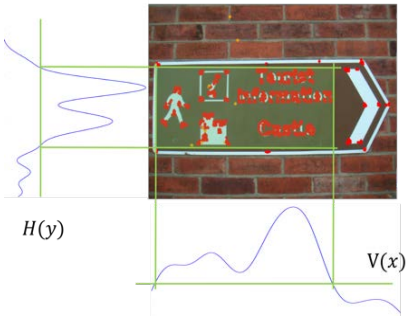
Method	Recall	Precision	Hmean
BUCT_YST	0.74	0.85	0.79
Our method	0.74	0.70	0.72

TABLE II. Performance Comparison for Taiwan Street Scene

Method	Recall	Precision	Hmean
Our method	0.79	0.81	0.80
BUCT_YST	0.35	0.34	0.34

IV. Conclusions

We have presented a novel text detection method. The method is based on the feature point extraction and the color distribution of the feature points. According to our experiments and analysis, the proposed method can be applied on signboards on street view images. The proposed method was also robust to various font sizes, font styles, and contrast levels. A known limitation of the current method is that the performance will be degraded if the text contains less feature points, such as junction, turning, and end points. These issues will be addressed in our future research.



(a)

(b)

(c)

Figure 6. (a) Vertical and horizontal projection (b) Text map image after morphological closing. (c) Minimum bounding boxes for the text area.

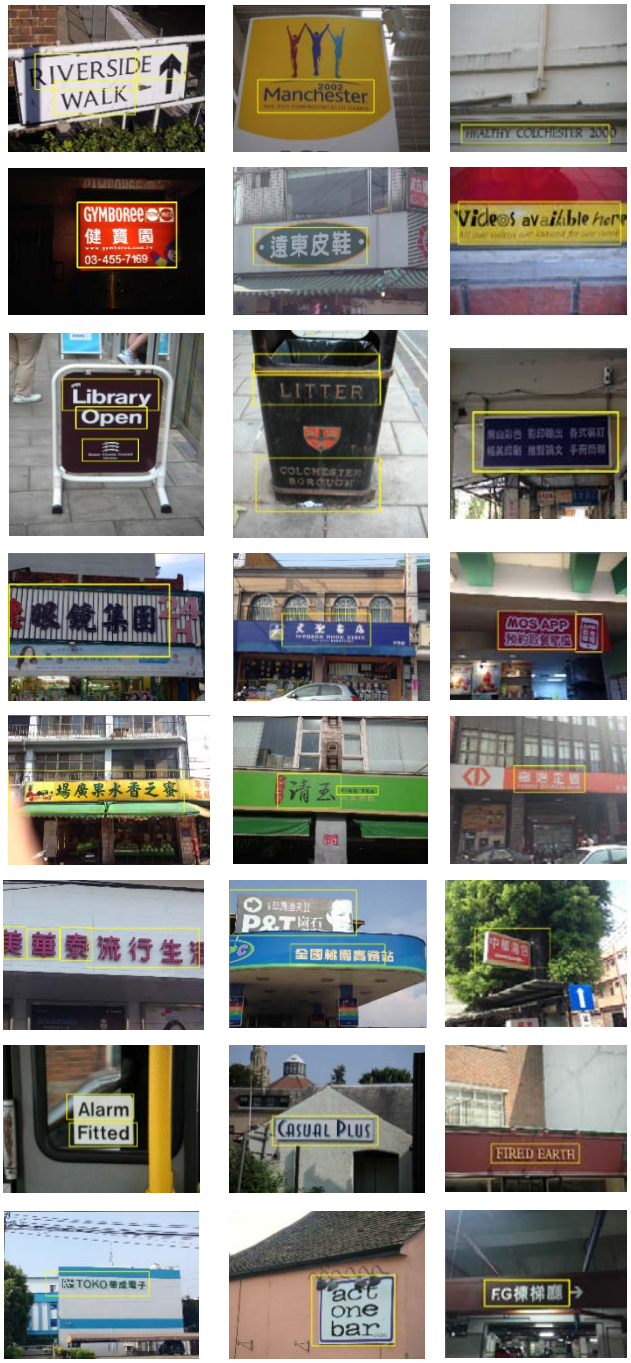


Figure 7. Examples of experimental results.

Acknowledgment

Thanks for the Ministry of Science and Technology (Grant number: MOST 105-2622-E-155-015 -CC3) for funding this work.

References

[1] S. A. Angadi and M. M. Kodabagi, "A texture based methodology for text region extraction from low resolution natural scene images," *International Journal of Image Processing*, vol. 3, pp. 121-128, 2010.

[2] K. I. Kim, K. Jung, and J. H. Kim, "Texture-based approach for text detection in images using support vector machines and continuously adaptive mean shift algorithm," *IEEE Transactions*

on Pattern Analysis and Machine Intelligence, vol. 25, pp. 1631-1639, 2003.
[3] X. Liu and J. Samarabandu, "Multiscale edge-based text extraction from complex images," in *Proc. IEEE International Conference on Multimedia and Expo*, 2006, pp. 1721-1724.
[4] X. Liu and J. Samarabandu, "An edge-based text region extraction algorithm for indoor mobile robot navigation," in *Proc. IEEE International Conference on Mechatronics and Automation*, 2005, vol. 2, pp. 701-706.
[5] H. Wang, "Automatic character location and segmentation in color scene images," in *Proc. IEEE International Conference on Image Analysis and Processing*, 2001, pp. 2-7.
[6] H. Wang and J. Kangas, "Character-like region verification for extracting text in scene images," in *Proc. 12th International Conference on Document Analysis and Recognition*, 2001, pp. 957-962.
[7] W. E. Wright, "Gravitational clustering," *Pattern Recognition*, vol. 9, pp. 151-166, 1977.
[8] J. Gomez, D. Dasgupta, and O. Nasraoui, "A new gravitational clustering algorithm," in *Proc. SIAM International Conference on Data Mining*, 2003, pp. 83-94.

Chien-Cheng Lee received the Ph.D. degree in electrical engineering from National Cheng Kung University, Tainan, Taiwan in 2003. Dr. Lee is currently an Assistant Professor in the Department of Communications Engineering, Yuan Ze University, Chungli, Taiwan. He has been a research visitor at the Telcordia Inc. (formerly Bellcore), NJ, USA, from Oct. 2007 to Jan. 2008. He is one of the guest editors for a special issue on Signal Processing for Applications in Healthcare Systems for EURASIP Journal on Advances in Signal Processing, 2008. His research interests include image processing, pattern recognition, and machine learning.

Shang-Fei Shen received his master degree in communications engineering from Yuan Ze University, Taoyuan, Taiwan, in 2015. His research interests include pattern recognition, image processing, and neural networks.