A Chinese Words Detection Method in Camera Based Images

Qingmin Chen, Yi Zhou, Kai Chen, Li Song, Xiaokang Yang
Institute of Image Communication and Information Processing, Shanghai Key Laboratory
Shanghai Jiao Tong University
Dongchuan Road 800, Shanghai, 200240, China
Email: qmchen2011@gmail.com, zy_21th@sjtu.edu.cn,
{ kchen, song_li, xkyang}@sjtu.edu.cn

Abstract. Text in camera based image carries important information for visual content understanding and retrieval. Locating text from complex background is a challenge task and complex structure of Chinese words make it harder. In this paper, we develop an application which is targeted toward smart phone with camera. Our goal is to enable the smart phone to detect Chinese words in nature scenes. First of all, we implemente Stroke Width Transform [1] (SWT for short) of images on smart phones. Consequently, we improve the filtering process of detected connected components. Finally, we propose grouping methods of structure analysis to detect Chinese word in nature scenes. We have implemented the algorithm and tested it. Experiments show that our Chinese word detection approach works well.

Keywords: camera, SWT, filtering, group, connected components

1. Introduction

Recently, there has been an increase use of smart phone's camera in capturing nature scene images and camera based image analysis becomes a hot research field [2]. Text in images includes useful information for the automatic annotation, indexing, visual content understanding and so on. Detecting text in images becomes an important step of many computer vision applications.

In this paper, we develop an application on smart phone to detect Chinese words and it is robust enough to text string with multiple sizes and colors, and arbitrary orientations in nature scene images with complex background. Figure 1 shows our flowchart of process.

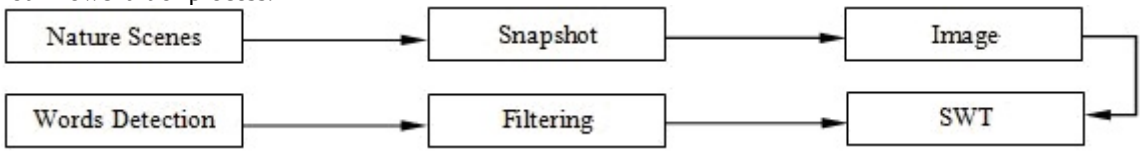
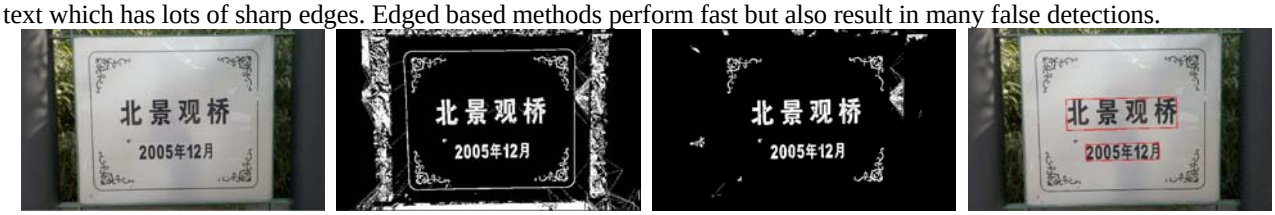


Fig.1 Flowchart of process

Previous work on text detection in images can be briefly classified into three domains: region-based [3], texture-based [4,5] and edge-based methods [6,7]. Region-based methods detect character as monochrome regions satisfying some similarities. One feature that separates text region from background is color similarity. However, the monochrome constraint can not be always satisfied. When it comes to complex background like nature scenes in which there are many foliages, the method is not Robust. Texture-based methods utilize texture features to determine whether a pixel or block of pixels belongs to text or not. Most texture based methods are robust ,but the high complexity of texture computation become drawback when most smart phones have limited memory. Some edge-based methods consider image block as text which has lots of sharp edges. Edged based methods perform fast but also result in many false detections.



(a) (b) (c) (d)

Fig.2 The detail of our method. (a) Input image capture by our smart phone. (b) The output of SWT. (c) After filtering, we prune out some connected components which are not text. (d) Implementing our Chinese word detection method, we locate the Chinese words with bounding boxs.

Recently, Epthtein *et al* [1] designed Stroke Width Transform to extract text characters with similar stroke width. It outperforms latest published algorithms. Benefiting from the work of Chucai Yi *et al* [10], we defined some constraints to group connected components and designed methods of structure analysis to detect Chinese word in nature scenes. The detail of our method is depicted in Figure 2.

The main contribution lies in two ways:

1) We give a detailed implementation of SWT of input images and improve the filtering process. For Chinese words, they have different morphology features compared with Latin words. The stroke width of Chinese words is closely related to the size and the shape tends to be square. We design corresponding filtering rules and filtering process is

improved.

2) We propose Chinese words grouping methods of structure analysis. Traditional grouping methods may fail to detected some Chinese words with complicated structures. Latin characters may align approximately horizontal, but for Chinese words, a single word may consist of a few " simple" charaters which align both horizontally and vertically under some constraints.

The rest of this paper is organized as follows. After a brief review of main point and drawback of the SWT and grouping method in Chucai Yi *et al* 's [10] work in Section 2, the detail of the SWT and our proposed grouping method is described in Section 3. The role of filtering is also detailed. The experiments and results analysis is shown in Section 4. A conclusion is drawn in Section 5.

2. Previous work

Epthtein *et al* [1]designed the Stroke Width Transform to recover the probable width of stroke. The Stroke Width Transform computes per pixel the most likely stroke width containing the pixel. The input of the SWT can be color image or gray image and the output of the SWT is an image of size equal to the input image. The pixel of the output image contains the likely stroke width. Then they group the pixels who have the similar stroke width. After the connected components are filtered which are obviously not text, methods of structure analysis of text strings are performed to detect text strings.

Epthtein *et al* [1] and Chucai Yi *et al* [10] designed two slightly different methods to group connected components to detect text lines. Both of them are designed to detect Latin words. They do not work well to detect Chinese words. Experimental results are shown in Section 4.

Though the mentioned grouping methods work well to detect Latin letters, it result in false detection of Chinese words because of Chinese words' complex structure. Chinese words have their very special structure which is quite different from Latin strings. Latin text strings may align approximately horizontal, but for Chinese word, a single word may consist of a few " simple" words which align both horizontally and vertically under some constraints. For example, the Chinese word " xiang" in Figure 3 can be divided into three Chinese characters: " mu", " mu" and " xin" in Figure 3. The first character "mu" in Figure 3 and the second character "mu" in Figure are in a horizontal distribution while the character " mu" and the character "xin" are in a approximate vertical distribution. Most of the existing grouping method focus on independent analysis of single character, and group letters into text lines. It fails to distinguish Chinese word from unwanted noises. To solve this problem, we propose two methods of structure analysis of Chinese words: intra - word grouping method and inter-word grouping method. Experiments show that our Chinese words detection approach works well.

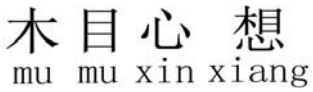
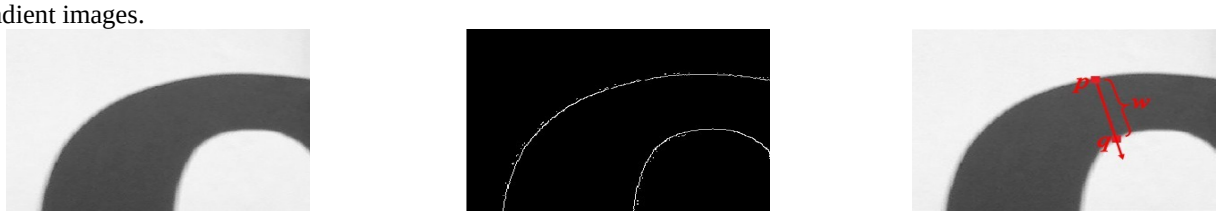


Fig.3 Example picture of Chinese characters.

3. The Detail of the Methods

A. Stroke Width Transform

We compute stroke width begins with gradient analysis [8, 9]. The gradient information of text is shown in Figure 4. To obtain the gradient image of input image, we respectively convolve the image with a horizonal and a vertical 3 × 3 Sobel kernel. We can get he gradient orientation of every pixel in image from the pixel at the same location of the gradient images.



(a) (b) (c)

Fig.4 The process of SWT. (a) A typical stroke of input image. (b) The edgemap of the stroke image. (c) ***p*** is on the edge of stroke. Finding another pixel ***q*** on the edge along the gradient direction of ***p***. The pixel couple is connected by a path. The value of pixels of the connecting path is assigned the stroke width ***w***.

The detail of the algorithm is clearly described below:

1) First of all, the initial value of pixel of output image is set to ∞. After that, we compute the edges in the image using Canny edge detector and the gradient images using Sobel edge detector.

2) For a pixel ***p*** locate on the edge of text and along the gradient orientation ***dp*** of the pixel ***p*** you find another pixel ***q***, a pixel couple is defined if ***dp*** is roughly opposite to *dq*.

3) The value of pixels in the connecting path between the pixel couple is set to width ||***p-q***|| unless it already has a lower value. repeat the above process.

4) Pass the connecting paths again, compute median SWT value of all pixels. Set the pixel whose value is above

median SWT to median SWT.

The computation is linear in the number of edge pixels in the image and its calculation is simple. It can satisfy the demand for smart phones with limited memory.

B. Filtering

The output of SWT is a image whose pixel contains the likely stroke width. We then modify the association rule of the connected domain labeling algorithm to group the neighbouring pixels if their value ratio does not exceed 3.0. After the the connected domains are labelled, we filter the domain which is obviously not text. For this reason we designed some rules which is more suitable for Chinese words than Latin characters. We compute the bounding box for each connected component for the convenience and correctness of the filtering. The centroid of the bounding box has the better representation of information such as position of the connected components than other information. The filtering process is performed to remove the connected components by stroke variance, size, aspect ratio and so on.

Firstly we compute the variance of stroke width within each component and discard the ones whose variance is too big. Experiments in the work[1] show that this rule suffice to distinguish between text and non-text like foliage.

Secondly, we compute the bounding rectangle for each connected components. For long and narrow components whose aspect ratio is not between 0.1 and 10, discard them.

Thirdly, we compute the bouding rectangle's area and the connected component's area. Take into consideration the condition that the above process will likely generate some random connected components consist of a few connecting paths whose bounding box is so big , then if the two-area ratio is too small, (for example, the ratio is lower than 0.1) we discard it.

Fourthly, for Chinese words, the stroke width is closely related to the size. If the width and the height are more than five times as long as the stroke width, the connected component will be discarded.

Lastly, connected component whose size is too big or too small is discarded.

After the filtering process, the connected components which are not discarded is considered as text.

C. Chinese word grouping

The last step is structural analysis of text strings which consists of two methods: intra-word detection method and inter-word method.

Assuming that a Chinese word has at least two " simple" words. We design two methods to locate Chinese words according to their similar stroke width and space distribution. We perform the intra-word process to detect single word followed by the inter-word process to detect text line which consists of at least two word.

The detail of grouping method is described below. A connected component is described by four metrics: height(.), width(.), centroid(.)(it is the centroid of the bounding box), mean(.)(the average of stroke width of a particular connected component) and D(.) to represent the distance between the centroids of adjacent connected component.

1) Intra-word Grouping method

Each connected component is considered a probable simple word and a Chinese word usually consists of at least two "simple" words. If a word has other words at adjacent positions and satisfy the constraints we defined, the intra-word grouping method will be performed. Here is the five constraints to determine whether the two connected components are grouped or not.

a. The average stroke width of the connected component is considered. The average stroke width ratio should fall into **1/T1** and **T1**.

b. We consider the two words align horizontally, if the difference between x-coordinates of the bounding box centroid is not greater than **T2** times the average stroke width.

c. We consider the two words align vertically, if the difference between y-coordinates of the bounding box centroid is not greater than **T3** times the average stroke width.

d. If the two words align horizontally, the difference between y-coordinates of the bounding box centroid is not greater than **T4** times the average stroke width.

e. If the two words align vertically, the difference between x-coordinates of the bounding box centroid is not greater than **T5** times the average stroke width.

We set **T1** = 0.8, **T2** = **T5** = 3.5 and **T4** = **T3** = 7. When two connected components satisfy the five constraints, we define a group **SG1** as the union of the two. When groups are created, merging process is performed. We merge two groups when the intersection of **SG1** and **SG2** contains at least one connected components. When no groups can be merged together, the intra-word grouping process is over. we consider the group created by the intra-word grouping process is a word. Before we start the inter-word grouping process, we then compute the bounding box of the word and reset the average stroke width.

2)Inter-word Grouping method

The inter-word grouping method is performed to locate text line. It is similar to the intra-word grouping method except that the constraints change. As is described above, after the inter-word grouping process, some connected components stay unchanged while some connected components are merged into a big one. Three constraints are described as follows:

a. The words in the same text line have similar average stroke width, so the average stroke width ratio should fall into **1/T6** and **T6**.

b. For the two words align horizontally, the difference between y-coordinates of the bounding box centroid should not be greater than **T7** times the height of the higher one.

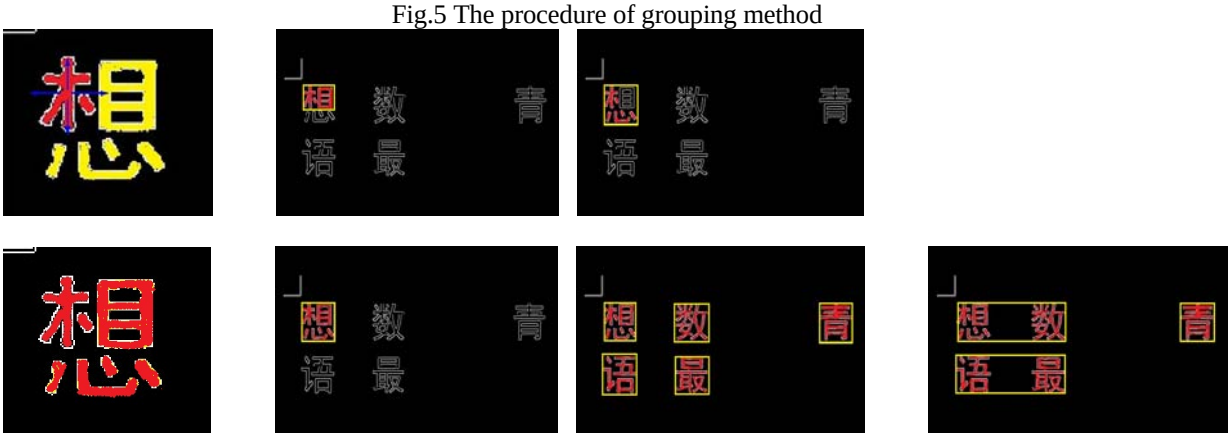
c. The distance between two adjacent words should be too far, so the difference between y-coordinates of the

bounding box centroids should be greater than **T8** times the width of the wider one.

We set **T6** = 0.8, **T7** = 0.5 and **T8** = 3. Then a merging process is performed. The merging process is similar with the above merging process. When there is no group has intersection, the inter-word grouping process is over. The connected components which stay unchanged are discarded. Figure 5 summaries the detail of our algorithm. Figure 6 shows the process of the grouping method.

```
0: Intra-word grouping process
1: S ← the set of connected components.
2: for each connected component C ∈ S
3:    for each connected component Ć ∈ S, and Ć ≠ C
4:       if T1 < mean(C)/ mean(Ć) < 1/T1
5:          if D(centroid(C).x - centroid(Ć).x) <
6:                          T2 * min{ mean(C), mean(Ć)}
7:                if D(centroid(C).y- centroid(Ć).y) <
8:                          T4 * min{ mean(C), mean(Ć)}
9:                   SG := Ć ∪ {C}
10:               endif
11:         else if D(centroid(C).x - centroid(Ć).x) <
12:                         T3 * min{ mean(C),mean(Ć)}
13:               if D(centroid(C).y - centroid(Ć).y) <
14:                         T5 * min{ mean(C),mean(Ć)}
15:                  SG := Ć ∪ {C}
16:               endif
17:      endif
18:   endfor
19: endfor
20: Repeat merging until no groups can be merged.
21:  for each group SG1
22:       for each group SG1 ≠ SG2
23:          if | SG1 ∩ SG2 | >= 1
24:             SG1:= SG1 ∪ SG2, SG2:= Ø,
25:          endif
26:       endfor
27:  endfor
28:
29: Inter-word grouping process
30: S ← the set of words created by intra-word grouping
31:      process and unchanged connected components
32: for each connected component C ∈ S
33:    for each connected component Ć ∈ S, and Ć ≠ C
34:       if T6 < mean(C)/ mean(Ć) < 1/T6
35:          & D(centroid(C).y - centroid(Ć).y) <
36:             T2*max{ height(C),height(Ć)}
37:          & D(centroid(C).x - centroid(Ć).x) <
38:                      T2*max{ width(C),width(Ć)}
39:             SG := Ć ∪ {C};
40:       endif
41:    endfor
42: endfor
43: Repeat the same merging process as the intra-word
44: process until no groups can be merged
```

Fig.5 The procedure of grouping method



(a) (b) (c)

Fig.6 The process of grouping. (a) Searching for adjacent connected components. (b) Intra-word merging process of "xiang" in Figure 3. (c) Two adjacent detected words merge into a text line. Two text lines and a word are detected.

The groups created by the intra-word grouping method is considered as input connected components of the inter-word grouping method. When the inter-word grouping process is finished, the groups is taken as a text line if they have no less than 2 connected components. If the group has a single connected component and it is the result of intra-word grouping method, it is taken as a single word.

4. Experiments and results

To testify our algorithm we ran our method on the public available dataset in [11] and our dataset of 78 images. The former one was used in two recent text detection competitions: ICDAR 2003 [12] and ICDAR 2005 [13]. The ICDAR dataset contains 258 images in the training set.The images' size vary from 307 × 93 to 1280 × 960 pixels and they are full-color. The other dataset contains 68 images. The images were taken by us using smart phone Milestone of Motorola. The images' size vary from 1073 × 583 to 2592 × 1456 and they are full-color.

The algorithm currently runs under 3 seconds on a 2592 × 1456 images, and less than 1 second on smaller images. The time consumed is short enough for smart phone applications.

In order to compare with the results in ICDAR 2003 database [13], the detected text lines have to be be broke into text words. While the problem in general does not require this step [1], we use another standard to judge the performance of our methods. Epthtein *et al* [1]test their algorithm on the ICDAR database and the result is: precision =

0.73, recall = 0.60.

There is no universal standard to judge the performance of all the methods of text location. We use two rates to show our method:

(i) Detection rate (D-rate):

$$D-rate=\frac{Number\_of\_Text\_Blocks\_Detected}{Number\_of\_Total\_Text\_Blocks}\times 100\% \quad (1)$$

(ii) False-alarm rate (F-rate):

$$F-rate=\frac{Number\_of\_False-alarm\_Blocks\_Detected}{Number\_fo\_Total\_Blocks\_Detected}\times 100\% \quad (2)$$

The two rates are computed on the two image dataset. The results on the ICDAR dataset are as follows: D-rate (average): 78.2%; F-rate (average) 46.6%. The results on our dataset are as follows: D-rate (average) 87.4%; F-rate (average): 38.8%. Some experimental results of our algorithm tested on the ICDAR dataset are shown in Figure 7. Some experimental results of our algorithm tested on our dataset are shown in Figure 8.

5. Conclusions

An application on smart phone to detect Chinese words has been proposed and its performance has been evaluated via several experiments. We use a distinctive local image operator SWT to represent features of the camera based image. Our grouping methods are robust to detect Chinese words in nature scenes. Low complexity of SWT computing is very important for smart phone application. The improved filtering process reduces the computation cost consequently. Our proposed grouping method is simple, steady and efficient to apply. And also we can scale the captured images' size to reduce computation cost to satisfy the demand of smart phones.

Though the application works efficiently to detect Chinese word, it can be further improved. The filtering constraints may result in false filtering and it needs improvement. More images will be tested to find the grouping method's drawback to get a better performance.

Fig.7 Experimental results tested on the ICDAR dataset

Fig.8 Example results of our method.

6. Acknowledgement

This work is supported by National Basic Research Program of China (2010CB731401 and 2010CB731406).

References

[1] B. Epshtein, E. Ofek, and Y. Wexler, "Detecting Text in Nature Scenes with Stroke Width Transform," In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2011.

[2] J. Liang, D. Doemann, and H.P. Li, "Camera-based analysis of text and documents: a survey," International Journal on Document Analysis and Recognition, vol. 6, no. 2-3, pp. 84-104, 2005.

[3] David Crandall, Sameer Antani and Rangachar Kasturi, " Extraction of Special Effects Caption Text Events Events From Digital Video," International Journal on Document Analysis and Recognition(IJDAR), vol. 5, pp. 138-157, 2003.

[4] Julinda Gllavata, Ralph Ewerth and Bernd Freisleben, "Text Detection in Images Based on Unsupervised Classification of High Frequency Wavelet Coefficients," Proceeding of 17th International Conference on Pattern Recognition (ICPR), vol. 1, pp. 425-428, 2004.

[5] Yangxing Liu, Satoshi Goto and Takeshi Ikenaga, " A contour-based Robust Algorithm for Text Detection in Color Images, " IEICE Transactions on Information and System,vol. E89-D, 1221-1230, 2006.

[6] P. Shivakumara, W. Huang and C.L. Tan. "Efficient Video Text Detection using Edge Features," ICPR 2008, December 8-11.

[7] P. Shivakumara, W. Huang and C.L. Tan. "An Efficient Edge based Technique for Text Detection in Video Frames, " The Eighth IAPR Workshop on Document Analysis Systems(DAS2008), Nara, Japan, pp. 307-314, 2008.

[8] P. Doucette, P. Agouris and A. Stefanidis, "Automated Road Extraction from High Resolution Multispectral Imagery," Photogrammetric Engineering & Remote Sensing, vol. 70. 12, pp. 1405-1416, 2004.
[9] C. Kirbas and F. Quek, "A review of vessel extraction techniques and algorithms", ACM Computing Surveys(CSUR), vol.36(2), pp. 81-121, 2004.
[10] Chucai. Yi and Yingli Tian, "Text String Detectin from Natural Scenes by Structure-based Partition and Grouping", IEEE Transactions on Image Processing. 2011.
[11] http://algoval.essex.ac.uk/icdar/Datasets.html.
[12] "ICDAR 2003 robust reading competitions", Proceedings of Seventh International Conference on Document Analysis and Recognition, 2003, pp. 682-68
[13] "ICDAR 2005 text locating competitions results", Eighth International Conference on Document Analysis and Recognition, 2005. Proceedings. pp. 80-84(1)