Improved Text-Detection Methods for a Camera-based Text Reading System for Blind Persons

Nobuo Ezaki[1], Kimiyasu Kiyota[2], Bui Truong Minh[3], Marius Bulacu[4] and Lambert Schomaker[4]
[1]*Toba National College of Maritime Technology, 517-8501 Mie, JAPAN*
[2]*Kumamoto National College of Technology, 861-1102 Kumamoto, JAPAN*
[3]*Tokyo Institute of Technology, 152-8550 Tokyo, JAPAN*
[4]*University of Groningen, Groningen, The Netherlands*
[1]*ezaki@toba-cmt.ac.jp,* [2]*kkiyota@tc.knct.ac.jp,* [4]*{M.Bulacu, L.Schomaker}@ai.rug.nl*

Abstract

Automatic text recognition from natural images receives a growing attention because of potential applications in image retrieval, robotics and intelligent transport system. Camera-based document analysis becomes a real possibility with the increasing resolution and availability of digital cameras.

Our research objective is a system that reads the text encountered in natural scenes with the aim to provide assistance to visually impaired persons. In the case of a blind person, finding the text region is the first important problem that must be addressed, because it cannot be assumed that the acquired image contains only characters.

In a previous paper [1], we propose four text-detection methods based on connected components. Finding small characters needed significant improvement. This paper describes a new text-detection method geared for small text characters. This method uses Fisher's Discriminant Rate (FDR) to decide whether an image area should be binarized using local or global thresholds. Fusing the new method with a previous morphology-based one yields improved results.

Using a controllable webcam and a laptop PC, we developed a prototype that works in real time. At first, our system tries to find in the image areas with small characters. Then it zooms into the found areas to retake higher resolution images necessary for character recognition. Going from this proof-of-concept to a complete system requires further research effort.

1. Introduction

The number of visually impaired persons is steadily increasing due to diabetes, eye diseases, traffic accidents, aging and other causes. There are about 200,000 persons with acquired blindness in Japan. Computer applications that provide support to the visually impaired persons have become an important theme. A navigation system for pedestrians using GPS on cellular phones is in current use. It can provide support to a blind person willing to go outside unaccompanied. However, reaching the exact destination in a city environment still requires additional information, like the text signs besides an office entrance. The GPS-based navigation system would be usefully complemented by a portable text reading system. More generally, when a visually impaired person is walking in a man-made environment, it is important to be able to acquire the text information which is present in the scene. For example, a 'stop' sign at a crossing without acoustic signal has an important meaning. As another example, if the signboard of a store can be read, the shopping wishes of the blind person can be satisfied easier. In a previous publication [1], we described the system design for a camera-based reading system that extracts text information from natural scene images and we evaluated the effectiveness of several text-detection methods on a dataset of static images. We found that the effectiveness of different methods strongly depends on character size. Since in natural scenes the observed characters may have widely different sizes, it is therefore difficult to extract all text areas from the image using only a single method. This would especially be the case for the real-world images acquired by a visually impaired person. In our previous study, the

weakest method regarded the extraction of image areas containing small text characters, less than 30 pixels in height or width. Small characters appear very frequently in natural scene images and detecting then is important. Because a first captured image does not always contain sufficiently clear characters, zooming into the candidate text area becomes a necessity. This requires the use of a controllable camera with pan-tilt-zoom capabilities.

In this paper we describe and evaluate an improved method for finding small text characters. We developed an actual prototype working in real time using a controllable webcam connected to a laptop PC. This paper describes also its performance.

2. System design

Figure 1 shows the general configuration of our proposed system. The building elements are a laptop PC, a web camera and a voice synthesizer. Zooming, pan-tilt motion and auto-focus are essential functions required for the web camera. The system enters in action whenever a user presses a button specially designed in the user interface. The camera, in principle placed on the user's shoulder, acquires an image of the scene. Then the search for text areas is performed using several different methods exploiting edge, color and morphological information. In the current paper, we focus on text-detection methods geared for small characters. If text areas are detected in the initial input image, the camera zooms-in to obtain more detailed images of each candidate text area. The higher resolution characters are then recognized and read out to the blind person via a voice synthesizer.

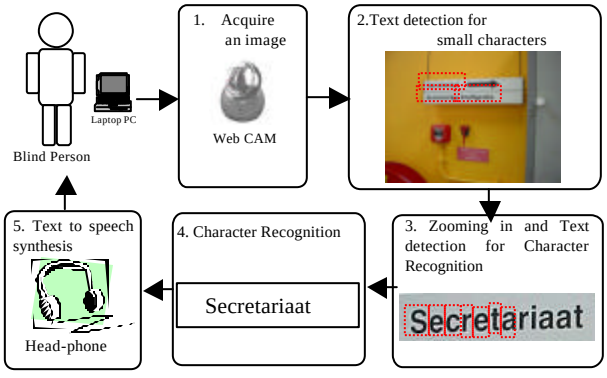


Fig.1 System Design

3. Text extraction method

3.1 Mathematical-morphology-based method

In previous work [1], we analyzed a character extraction method for small character (less than 32x32 pixels) based on mathematical morphology operations.

We used a modified top-hat processing capable to extract text areas based on the property that the text characters are very thin structures vulnerable to erosion.

The method requires that the text pixels have intensity values sufficiently different from the background, a global threshold being used for binarizing the image prior to connected component extraction. The method encounters difficulties for text with low contrast when a local threshold would be need. The new method we present in the next section of the paper addresses this problem.

3.2 FDR-based method

We propose a new extraction method that combines global and local image binarizations. Our technique is essentially based on Otsu's binarization method and it is applied in the same way in all three-color channels. The input image is divided in non-overlapping tiles of in 32x32 pixels. For every such tile, Fisher's Discriminant Rate (FDR) is computed from the histogram of the each color channel using equation (1):

$$FDR = s_B^2 / s_w^2 \qquad (1)$$

In equation (2), s_W^2 is within-class variance of the each color channel histogram. s_B^2 is between-class variance (equation (3)). Here s_1^2 and s_2^2 are variance of each class, w_1 and w_2 are probabilities of class occurrence, m_1 and m_2 are average of pixel values.

$$s_W^2 = w_1 s_1^2 + w_2 s_2^2 \qquad (2)$$

$$s_B^2 = w_1 w_2 (m_1 - m_2)^2 \qquad (3)$$

If characters are present in a tile (Fig.2), then the local histogram has two peaks and this is reflected in a high value for the FDR. For quasi-uniform tiles (Fig.3(a)) the histogram has one peak and the value of the FDR is small. For more complex areas (Fig.3(b)) the histogram is dispersed resulting in higher FDR values, but which are still usually lower than in the case of text areas.

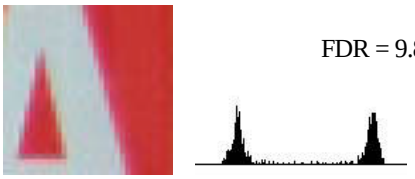


Fig.2 Character Area and Histogram

(a) Uniform Area (ex. Background)

(b) Complex Area

Fig.3 Examples of non Character Area

The FDR therefore can be used for detecting the image tiles with a bimodal gray-level histogram. For image tiles with high FDR values, the local Otsu threshold is used for binarizing the image. For tiles with low FDR values, the global Otsu threshold is used instead. A threshold value of 5.0 for the FDR was chosen through experimentation. After binarizing all three-color channels independently, each pixel can only have $2^3 = 8$ possible combinations of color values, the color space being therefore reduced to 8 colors. We separate the 8 binary images and then we extract connected-components on each one independently.

(a)

(b)

(c)

(a) Original Image
(b) 8-colored image by FDR method
(c) Extracted candidate character areas

Fig.4 Example of FDR based method

Fig. 4 shows example image of character extraction by FDR method. You can see an 8-colored image that some tiles are binarized by local threshold in Fig4 (b).

3.3 Connected-component extraction and selection

Up to this point, the proposed methods are very general in nature and not very specific to text detection. As expected, many of the extracted CoCos do not actually contain text characters. At this point simple rules are used to filter out the false detections. We impose constraints on the aspect ratio and area to decrease the number of non-character candidates. Fig. 5 gives an overview of the CoCo selection rules used in our method. *Wi* and *Hi* are the width and height of an extracted CoCo, Δx and Δy are the distances between their centers of gravity. Aspect ratio is computed as width / height. An important observation is that, generally, text characters do not appear alone, but together with other characters of similar dimensions and usually regularly placed in a horizontal string. We use selection rules based on the relative placement of CoCos to further eliminate from all the detected CoCos those that do not actually correspond to text characters.

The system goes through all combinations of two CoCos and only those complying with all the selection rules succeed in becoming part of the final proposed text region.

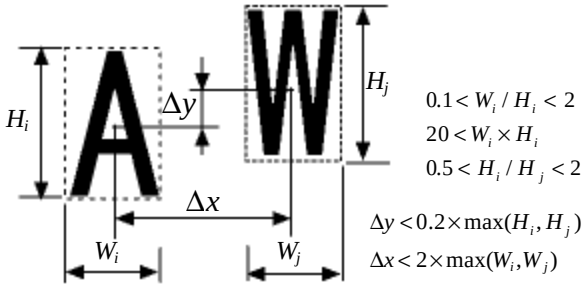


Fig. 5 Connecting Component Rules

4.Evaluation Experiments

For evaluating the performance of the proposed methods, we used the dataset made available with the occasion of the *ICDAR 2003 Robust Reading Competition* [5]. The images are organized in three sections: Sample, Trial and Competition. Only the first two are publicly available, the third set of images being kept separate by the competition organizers to have a completely objective evaluation. The Trial directory has two subdirectories: *Trial-Train* and *Trial-Test*. The

Trial-Train images should be used to train and tune the algorithms. As we do not use machine learning in our approach, we included all the images in *Trial-Test* and *Trial-Train* for evaluation. This difficult dataset contains a total of 508 realistic images with textual content.

Our evaluation method is inspired from the one used in the ICDAR2003 competition, but it is much simpler. The ICDAR2003 evaluation computes the notions of precision and recall using words as detection instances. At the moment, our system does not have layout-analysis and/or character recognition modules that would allow finding words in the detected regions. We compute precision and recall using image areas expressed in terms of number of pixels.

Precision p is defined as the number of correct estimates C divided by the total number of estimates E. Recall r is defined as the number of correct estimates C divided by the total number of targets T.

$$p = c/|E|,\ r = c/|T|$$

E is the area proposed by our algorithm, T is the manually labeled text area and C is their intersection. We then compute the average precision and recall aver all the images in the dataset. There is usually a trade-off between precision and recall for a given algorithm. It is therefore necessary to combine them into a single final measure of quality f:

$$f = (\alpha / p + (1-\alpha)/r)^{-1}$$

The parameter α was set to 0.5, giving equal weights to precision and recall in the combined measure f.

4.1 Evaluation of Extraction method for Small Characters

For evaluating our FDR-based method for detecting small characters, we separated from the ICDAR dataset 75 images containing characters smaller than 30 pixels in height. Table 1 shows our experimental results. The newly proposed method yields higher precision and lower recall than the conventional method based on mathematical morphology.

Table 1 Results of Small Character Data set

	p	*R*	*f*
Proposed Method Classification by FDR	0.51	0.31	0.39
Conventional Method Morphological operation	0.47	0.35	0.40
ORing of 2 method	0.46	0.50	0.48

However, by ORing the two methods, a high recall value is achieved with a significant improvement on the combined f value as well. This means that each method extracts somewhat complementary regions and an OR fusion is beneficial for finding as much as possible of the small character areas present in the image.

4.2 Evaluation of Extraction method for Large Characters

Table 2 shows result of our extraction method for large characters [1]. The Sobel edge-based text detection method obtained top overall performance.

Table 2 Results of All Images

	p	*r*	*f*
Sobel edge-based	0.60	0.64	0.62
Reverse	0.62	0.39	0.50
8-colored	0.56	0.43	0.49
ORing of 3 method	0.51	0.73	0.62

5. Description of a working prototype

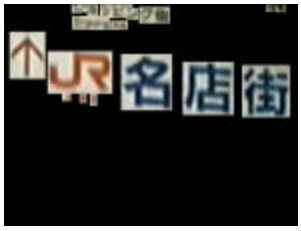
In order to directly text our methods in a more dynamic setting, we have built a working system. A Logitech QVR-1 camera with zooming, pan-tilt motion and auto-focus functions was coupled to a laptop PC.

Fig. 6 shows example of captured images by the prototype system. When a user wants to obtain the textual information present in the scene, he/she presses the capture button. The camera, placed on his/her shoulder, will automatically search for text areas. Our prototype system employs several text-detection methods in order to minimize the miss rate. This prototype system works in real time. The FDR based method is applied to first captured images (Fig.6 (a)) to find small text areas, this method works within 200ms (Intel Pentium4 1.4GHz). Since the images might contain large character, we also use the Sobel method. The system memorizes target positions of grouped candidate areas. After zooming in, Sobel method is applied to extract large characters for character recognition. Image processing time by Sobel method per frame is within about 300ms. The performance of our system is not perfect and it is very much in line with the reported precision and recall values in table 1 and table 2. In principle, it is a natural job for the character recognizer to reject many of the false text detections based on its knowledge of character shape in a complete system.

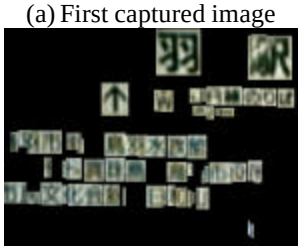
In addition, the methods described in this paper are used for detecting small characters. Up to the present, the detection/pan-tilt-zoom functionality was tested in

the prototype. Character recognition has not been implemented yet.

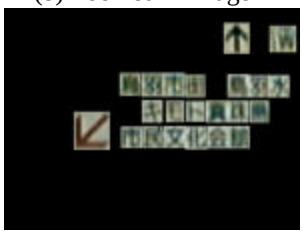
(a) First captured image



(b) Zoomed in image 1



(c) Zoomed in image 2



(d) Zoomed in image 3

Fig. 6 Captured image by the prototype system

6. Conclusion and Future works

In this paper, we presented our research efforts aiming at developing a camera-based text reading system for visually impaired persons. We implemented and evaluated a new text-detection method for image areas containing small characters. The OR fusion between the new method and a previous morphology-based method achieves high recall rates and an *f* value of 0.50.

We built a working prototype with pan-tilt-zoom functionality. However, current results are not enough for practical use. Future work will focus on new methods for extracting text characters with higher accuracy and on the development of a full demonstration system.

References

[1] N.Ezaki, M.Bulacu and L.Schomaker, "Text Detection from Natural Scene Images; Towards a System for Visually Impaired Persons", Proc. 17th Int. Conf. on Pattern Recognition, 2004, pp.683-686.

[2] D. Doermann, J. Liang, and H. Li, "Progress in Camera-Based Document Images Analysis", *Proc.of the ICDAR*, 2003, pp. 606-616.

[3] N. Otsu, "A Threshold Selection Method from Gray-Level Histogram", *IEEE Trans. Systems, Man and Cybernetics*, Vol. 9, 1979, pp. 62-69.

[4] Fisher R. A. "The use of multiple measurements in taxonomic problems. Amals. Eugenics", 7, 1936,179-188.

[5] S.M. Lucas, A. Panaretos, L. Sosa, A. Tang, S. Wong, and R. Young, "ICDAR 2003 Robust Reading Competitions", *Proc.of the ICDAR*, 2003, pp. 682-687.

[6] T. Yamaguchi, Y. Nakano, M. Maruyama, H. Miyao and T.Hananoi, "Digit Classification on Signboards for Telephone Number Recognition", *Proc.of the ICDAR*, 2003, pp.359-363.

[7] K.Matsuo, K.Ueda and M.Umeda, "Extraction of Character String from Scene Image by Binarizing Local Target Area", *T-IEE Japan*, Vol. 122-C(2), 2002, pp.232-241.

[8] Y. Liu, T. Yamamura, N. Ohnishi and N. Sugie, "Extraction of Character String Regions from a Scene Image", *IEICE Japan*, D-II, Vol. J81, No.4, 1998, pp.641-650.

[9] L. Gu, N. Tanaka, T. Kaneko and R.M. Haralick, "The Extraction of Characters from Cover Images Using Mathematical Morphology", *IEICE Japan*, D-II, Vol. J80, No.10, 1997, pp. 2696-2704.

[10] J. Yang, J. Gao, Y. Zhang, X. Chen and A. Waibel, "An Automatic Sign Recognition and Translation System", *Proceedings of the Workshop on Perceptive User Interfaces (PUI'01)*, 2001, pp. 1-8.

[11] A. Zandifar, R. Duraiswami, A. Chahine, and L. Davis, "A Video Based Interface to Textual Information for the Visually Impaired", *IEEE 4th ICMI*, 2002, pp.325-330.

[12] K.C.Kim, H.R.Byun, Y.J.Song Y.W.Choi, S.Y.Chi, K.K.Kim and Y.K.Chung, "Scene Text Extraction in Natural Scene Images by Hierarchical Classifiers", *Proc. 17th Int. Conf. on Pattern Recognition*, 2004, pp.679-682.

[13] S.Saitoh, H.Goto and H.Kobayashi, "Analysis and comparison of frequency features for scene text detection", Technical Report of IEICE, PRMU2004-128, pp31-36.