Text Detection from Natural Scene Images:
Towards a System for Visually Impaired Persons

Nobuo Ezaki
Information and Control Engineering Department
Toba National College of Maritime Technology, Japan
ezaki@toba-cmt.ac.jp

Marius Bulacu   Lambert Schomaker
AI Institute
Groningen University, The Netherlands
(bulacu, schomaker)@ai.rug.nl

Abstract

We propose a system that reads the text encountered in natural scenes with the aim to provide assistance to the visually impaired persons. This paper describes the system design and evaluates several character extraction methods. 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. However, 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. 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. In the present paper, we propose four character-extraction methods based on connected components. We tested the effectiveness of our methods on the ICDAR 2003 Robust Reading Competition data. The performance of the different methods depends on character size. In the data, bigger characters are more prevalent and the most effective extraction method proves to be the sequence: Sobel edge detection, Otsu binarization, connected component extraction and rule-based connected component filtering.

1. Introduction

Every year, the number of visually impaired persons is increasing due to eye diseases, diabetes, traffic accidents and other causes. There are about 200,000 persons with acquired blindness in Japan. Therefore computer applications that provide support to the visually impaired persons have become an important theme. We have already developed a pen-based character input system for blind persons using a PDA [2]. On this system, people with acquired blindness remember the shape and writing order of Japanese characters and they can use this system as a notepad and as an E-mail terminal anytime, anywhere. This application essentially works as communication tool. However, such a device does not solve all of the problems encountered by a blind person willing to go outside unaccompanied.

When a visually impaired person is walking around, it is important to get text information which is present in the scene. For example, a 'stop' sign at a crossing without acoustic signal has an important meaning. In general, way finding into a man-made environment is helped considerably by the ability to read signs. As an example, if the signboard of a store can be read, the shopping wishes of the blind person can be satisfied easier.

The research on text extraction from natural scene images has been growing recently [1]. Many methods have been proposed based on edge detection [8], binarization [6], spatial-frequency image analysis [4] and mathematical morphology operations [3]. Yang has proposed a sign recognition and translation system for tourists [9]: characters are extracted from images with Chinese sign boards and translated to English. There are also other parallel research efforts to develop a scene-text reading system for the visually impaired [10]. All these systems make evident that the text areas cannot be perfectly extracted from the image because natural scenes consist of complex objects, sometimes highly textured, buildings, trees, window frames and so on, giving rise to false text detection and misses. The first step in developing our text reading system is to address the problem of text detection in natural scene images. In this paper, we describe the system design and propose four text extraction methods based on connected components.

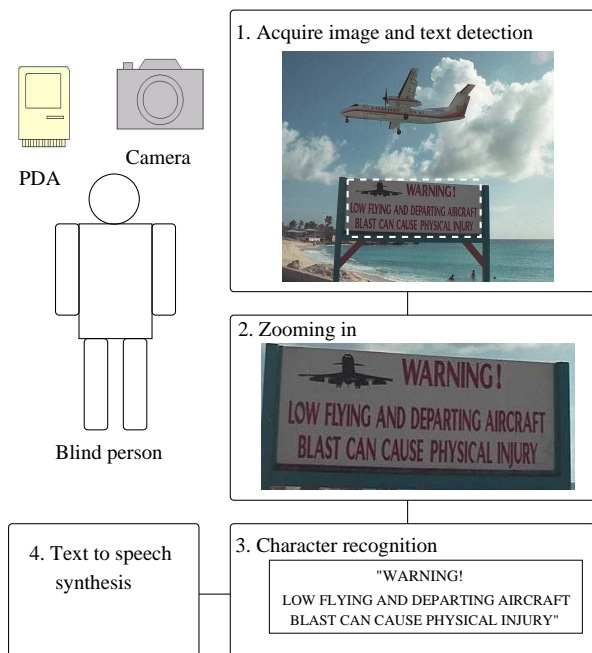Most studies are based on a single method for text detec-

Figure 1. System configuration ('walk-around mode')

tion. 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 is especially the case for the real-world images acquired by a visually impaired person. Under the envisaged usage conditions, the camera attitude will be much less constrained than is the case in current benchmark databases. We test the accuracy of the proposed character extraction methods on a newly available benchmark dataset assembled with the occasion of the IC-DAR 2003 Robust Reading Competition. We also evaluate how the individual methods can be combined for improving performance.

2. System design

Figure 1 shows the general configuration of our proposed system. The building elements are the PDA, the CCD-camera and the voice synthesizer. Zooming, pan-tilt motion and auto-focus are essential functions required for the CCD-camera.

Locating scene text involves two scenarios. First, in the 'walk-around mode', the camera which is placed on the user's shoulder acquires an image of the scene automatically and then the search for text areas is performed using methods geared for small characters. If an area is de-

tected, the camera zooms in to obtain a more detailed image on which extraction methods for large characters are used. These higher resolution characters are then recognized and read out to the blind person via a voice synthesizer. Of course, a gaze stabilization function is required in this mode, such that the system does not lose the target candidate character area while the user is walking. In this paper, however, we assume that the user is standing still when the images are captured.

In a second mode, the system is used for reading a restaurant menu or a book cover. In this scenario, the user can guess where the text is approximately and he/she can use the camera as a hand scanner. In this case, image resolution should need to be higher than in the 'walk-around mode' because it is expected that the images will contain many characters.

3. Extraction of small characters using mathematical-morphology operators

The first method we propose targets the small characters (less than about 30 pixels in height) and it is based on mathematical morphology operations. We use a modified top-hat processing [3]. In general, top-hat contrast enhancement is performed by calculating the difference between the original image and the image obtained after applying the opening operation on the original image. As a consequence, the top-hat operation is applicable when the pixels of the text characters have intensity values which are sufficiently different from the background. For instance, Gu et al. [3] use the difference between closing operation and the original image for text detection when character pixels have lower intensity values than the background (for light text on a dark background). This method is very effective, however it become computationally expensive if a large filter is used in order to extract large characters. We developed an invariant method applicable to small characters. We use a disk filter with a radius of 3 pixels and we take the difference between the closing image and the opening image. The filtered images are binarized and then connected components (CoCos) are extracted (Fig. 2b). This method detects connected text areas containing several small characters. As western text consists of strings of characters that are usually horizontally placed, we take horizontally long areas $(1 < width/height < 25)$ from the output image as the final candidate text regions (Fig. 2c).
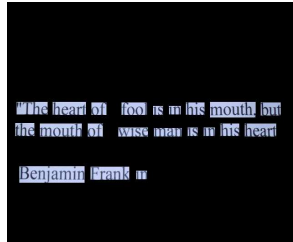
4. Three extraction methods for large characters

We propose three extraction methods for large characters (more than about 30 pixels in height). The first two
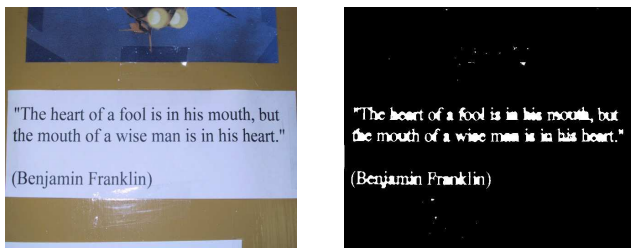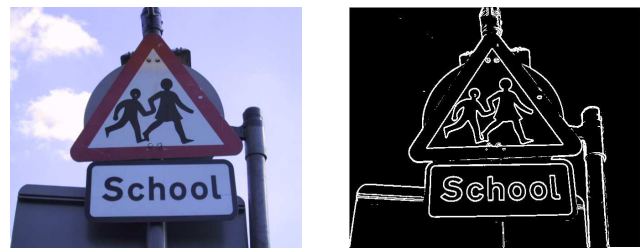
(a)

(b)

(c)

Figure 2. Extraction of small characters using morphological operations: a) original image, b) difference between closing and opening, c) extracted characters are found by a mask operator between the original image and the bounding rectangles of connected components in (b).

(a)

(b)

(c)

(d)

Figure 3. Example of an image with medium-size characters: a) original image, b) edge image, c) reverse edge image, d) 8-color image.

are based on Sobel edge detection and the third is based on RGB color information. In the overall system, these methods should be used after zooming into the areas initially found by the morphological operations. Each method extracts connected components that represent candidate text areas. Decision rules based on the sizes and relative positioning of these areas are afterwards used to prune the number of possibilities and reduce the large number of false hits.
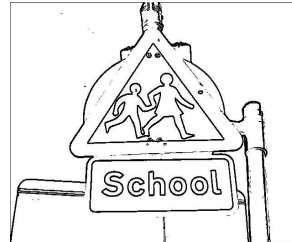
4.1. Character extraction from the edge image

In this method, Sobel edge detection is applied on each color channel of the RGB image. The three edge images are then combined into a single output image by taking the maximum of the three edge values corresponding to each pixel. The output image is binarized using Otsu's method [7] and finally CoCos are extracted.

This method will fail when the edges of several characters are lumped together into a single large CoCo that is eliminated by the selection rules. This often happens when the text characters are close to each other or when the background is not uniform.

4.2. Character extraction from the reverse edge image

This method is complementary to the previous one; the binary image is reversed before connected component extraction. It will be effective only when characters are surrounded by connected edges and the inner ink area is not broken (as in the case of boldface characters).

4.3. Color-based character extraction

The three methods proposed until now use morphological and edge information for text detection. However, color information is also important, because, usually, related characters in a text have almost the same color for a given instance encountered in the scene. The first step is to simplify the color space and we reduce it to 8 colors by the following procedure. We are applying Otsu binarization independently on the three RGB color channels. Each pixel can now have only $2^3 = 8$ possible combinations of color values. We separate the 8 binary images, then we extract and select CoCos on each one independently.
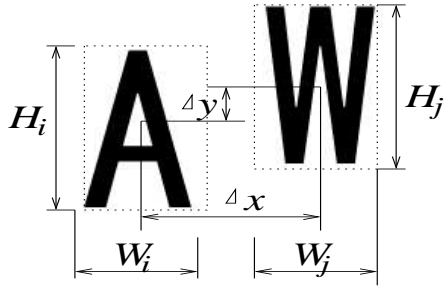
Figure 4. Character strings and rules.

Figure 5. Final result for the example given in fig. 3.

4.4. Connected-component selection rules

It can be noticed that, up to now, the proposed methods are very general in nature and not 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 size to decrease the number of non-character candidates. In Fig. 4, W_i and H_i are the width and height of an extracted area; Δx and Δy are the distances between the centers of gravity of each area. 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 the following rules to further eliminate from all the detected CoCos those that do not actually correspond to text characters:

$$0.1 < \frac{W_i}{H_i} < 2 \tag{1}$$

$$50 < W_i\, H_i \tag{2}$$

$$0.5 < \frac{H_i}{H_j} < 2 \tag{3}$$

$$\Delta y < 0.2\, max(H_i, H_j) \tag{4}$$

$$\Delta x < 2\, max(W_i, W_j) \tag{5}$$

The system goes through all combinations of two CoCos and only those complying with all the selection rules succeed to the final proposed text region (see fig. 5).

5. Evaluation experiment

For evaluating the performance of the proposed text detection 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 text detection methods, we included all the images in Trial-Test and Trial-Train for evaluation. This difficult dataset contains a total of 504 realistic images with textual content.

We used a similar evaluation method as that of the ICDAR2003 competition. It is based on the notions of precision and recall. Precision p is defined as the number of correct estimates C divided by the total number of estimates E:

$$p = \frac{C}{E} \tag{6}$$

Recall r is defined as the number of correct estimates C divided by the total number of targets T:

$$r = \frac{C}{T} \tag{7}$$

For a given image, we calculate precision and recall as the ratio between two image areas (expressed in terms of number of pixels). 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:

$$f = \frac{1}{\alpha/p + (1-\alpha)/r} \tag{8}$$

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

Table 1. Results for the individual text extraction methods.

Method	Precision	Recall	f
Edge (E)	60%	64%	62%
Edge reverse (R)	62%	39%	50%
8 colors (8)	56%	43%	49%
Morphology (M)	41%	16%	28%

Table 2. Results obtained after fusing methods using OR.

Method	Precision	Recall	f
E + 8	54%	69%	62%
E + R	56%	70%	63%
E + M	55%	68%	62%
E + R + 8	51%	73%	62%
E + R + 8 + M	48%	76%	62%

Our results on the ICDAR 2003 dataset are shown in Table 1. The edge-based text detection method obtained top overall performance. In this context, we note that, at IC-DAR 2003 [5], the results for the winner of the competition were precision = 55%, recall = 46% and f = 50%.

The morphological method did not obtain good overall results because the dataset contains relative large text characters. Consequently, we selected, from the ICDAR 2003 dataset, a group of 55 images that contain only small characters. We evaluated the efficacy of the morphological method on these images and obtained precision = 38%, recall = 55% and f = 47%. We tested also the edge based method on these images and obtained precision = 26%, recall = 48% and f = 37%. The morphological method seems to be more effective for small characters.

Table 2 shows the results obtained by combining methods. Fusion is performed by ORing the results of the individual methods. The increase in recall is outbalanced by the decrease in precision. However, for the same f value, the method with the highest recall rate is preferable. 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.

6. Conclusion

In this paper, we presented the design of a scene-text detection module within a reading system for visually impaired persons. As the first step in the development of this system, four connected-component-based methods for text detection have been implemented and evaluated. The most effective proves to be the sequence: Sobel edge detection, Otsu binarization, connected-component extraction and rule-based connected-component selection. A high recall rate can be achieved by collecting all the candidate text areas proposed by the four individual methods. However, current results are not enough for practical use. Future work will focus on new methods for extracting small text characters with higher accuracy.

References

[1] D. Doermann, J. Liang, and H. Li. Progress in camera-based document image analysis. In *Proc. of 7th Int. Conf. on Document Analysis and Recognition (ICDAR 2003)*, volume I, pages 606–616, Edinburgh, Scotland, 3-6 August 2003. IEEE Press.

[2] N. Ezaki, K. Kiyota, H. Takizawa, and S. Yamamoto. Pen-based ubiquitous computing system for visually impaired person. In *Human-computer Interaction, Theory and Practice (Part II)*, volume 2, pages 48–52, Crete, Greece, 22-27 June 2003. Lawrence Erlbaum Associates, Inc.

[3] L. Gu, N. Tanaka, T. Kaneko, and R. Haralick. The extraction of characters from cover images using mathematical morphology. *Transaction of The Institute of Electronics, Information and Communication Engineers of Japan*, J80-D-II(10):2696–2704, October 1997.

[4] Y. Liu, T. Yamamura, N. Ohnishi, and N. Sugie. Extraction of character string regions from a scene image. *Transaction of The Institute of Electronics, Information and Communication Engineers of Japan*, J81-D-II(4):641–650, April 1998.

[5] S. M. Lucas, A. Panaretos, L. Sosa, A. Tang, S. Wong, and R. Young. Icdar 2003 robust reading competitions. In *Proc. of 7th Int. Conf. on Document Analysis and Recognition (ICDAR 2003)*, volume II, pages 682–687, Edinburgh, Scotland, 3-6 August 2003. IEEE Press.
http://algoval.essex.ac.uk/icdar/RobustReading.html .

[6] K. Matsuo, K. Ueda, and U. Michio. Extraction of character string from scene image by binarizing local target area. *Transaction of The Institute of Electrical Engineers of Japan*, 122-C(2):232–241, February 2002.

[7] N. Otsu. A threshold selection method from gray-level histogram. *IEEE Trans. Systems, Man and Cybernetics*, 9:62–69, 1979.

[8] T. Yamaguchi, Y. Nakano, M. Maruyama, H. Miyao, and T. Hananoi. Digit classification on signboards for telephone number recognition. In *Proc. of 7th Int. Conf. on Document Analysis and Recognition (ICDAR 2003)*, volume I, pages 359–363, Edinburgh, Scotland, 3-6 August 2003. IEEE Press.

[9] J. Yang, J. Gao, Y. Zang, X. Chen, and A. Waibel. An automatic sign recognition and translation system. In *Proceedings of the Workshop on Perceptive User Interfaces (PUI'01)*, November 2001.

[10] A. Zandifar, R. Duraiswami, A. Chahine, and L. Davis. A video based interface to textual information for the visually impaired. In *Proc. of 4th Int. Conf. on Multimodal Interfaces (ICMI 2002)*, pages 325–330, Pittsburgh, USA, 14-16 October 2002.