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


## 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 [1]. 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 in the last several years [2]. Many methods have been proposed based on edge detection [3], binarization [4], spatial-frequency image analysis [5] and mathematical morphology operations [6]. Yang has proposed a sign recognition and translation system for tourists [7]: 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 [8]. All these systems make evident that the text areas cannot be perfectly extracted from the image because natural scenes consists 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 detection. 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 ICDAR 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 CDD-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 detected, 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.

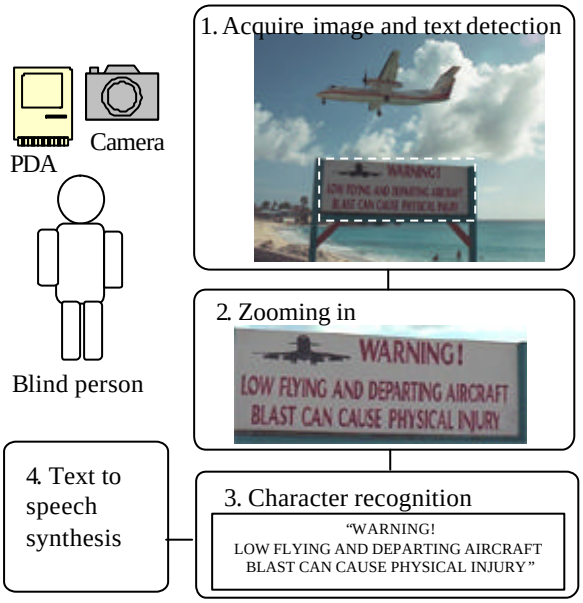
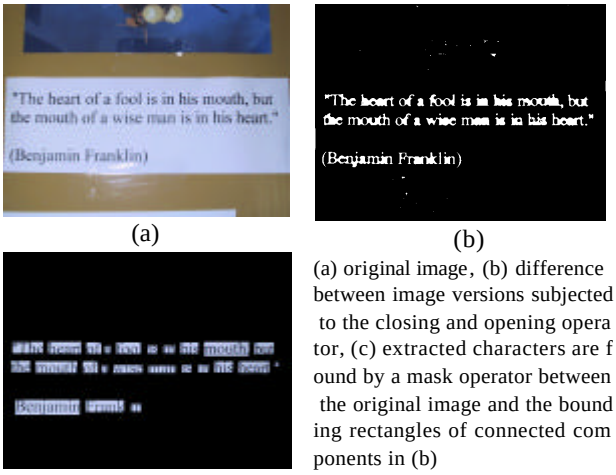



Figure1 System configuration (walk-around mode)

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 operations

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 [6]. In general, top-hat contrast enhancement is performed by calculating the difference between the original image and the image obtained after applying the opening image on the original image. As a consequence, the top-hat operation is applicable when the pixels of the text characters have intension values which are sufficiently different from the background. For instance, Gu [6] uses the difference between closing operation and the original image for text detection when character pixels have lower 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).




(a) (b)

(c)

(a) original image, (b) difference between image versions subjected to the closing and opening operator, (c) extracted characters are found by a mask operator between the original image and the bounding rectangles of connected components in (b)

Figure2 Extraction of small character text using morphological operation

## 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 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 [9] and finally CoCos are extracted.

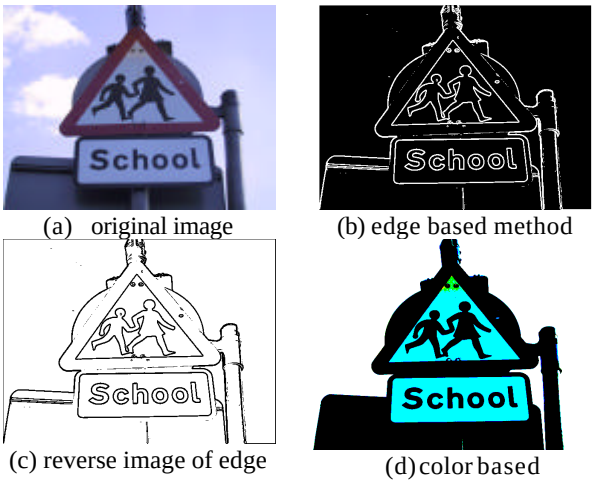
This method will fail when the edges of several characters are lumped together into a single large CoCo that is eliminated by the selection rules. This often happens when the text characters are close to each other or when the background is not uniform.

### 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.




(a) original image (b) edge based method

(c) reverse image of edge (d) color based

Figure 3 Example of an image with medium-size character

### 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; $\Delta x$ and $\Delta y$ 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 (Fig.4):

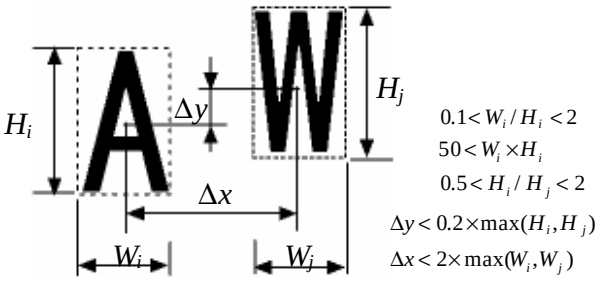



Figure 4 Character strings and rules

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


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

## 5. Evaluation experiment

For evaluating the performance of the proposed methods, we used the dataset made available with the occasion of the *ICDAR 2003 Robust Reading Competition* [10]. 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 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 $E$:

$$p = c / |E|$$

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

$$r = c / |T|$$

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 $f$:

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

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

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 ICDAR 2003, the results for the winner of the competition were precision = 0.55, recall = 0.46 and $f$ = 0.50. 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 = 0.38, recall = 0.55 and $f$ = 0.47. We tested also the edge based method on these images and obtained precision = 0.26, recall = 0.48 and $f$ = 0.37. The morphological method seems to be more effective for small characters.

Table 1 Results for the individual text extraction methods. (Proportion correct characters)

| | Precision | Recall | $f$ |
|---|---|---|---|
| Edge(E) | 0.60 | 0.64 | 0.62 |
| Reverse(R) | 0.62 | 0.39 | 0.50 |
| 8 color(8) | 0.56 | 0.43 | 0.49 |
| Morphology(M) | 0.41 | 0.16 | 0.28 |

Table 2 Results obtained after fusing methods using OR. (Proportion correct characters)

| | Precision | Recall | $f$ |
|---|---|---|---|
| E+8 | 0.54 | 0.69 | 0.62 |
| R+E | 0.56 | 0.70 | 0.63 |
| E+M | 0.55 | 0.68 | 0.62 |
| R+E+8 | 0.51 | 0.73 | 0.62 |
| R+E+8+M | 0.48 | 0.76 | 0.62 |

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.

## 7. References


[1] N. Ezaki, K. Kiyota, H. Takizawa and S. Yamamoto, "Pen-based Ubiquitous Computing System for Visually Impaired Person", human-computer interaction, Theory and Practice (PartII), Volume2, 2003, pp.48-52.

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

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

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

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

[10] 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.