Towards Text Recognition in Natural Scene Images

Basilios Gatos, Ioannis Pratikakis and Stavros Perantonis

Computational Intelligence Laboratory, Institute of Informatics and Telecommunications, NCSR "DEMOKRITOS" Athens 153 10, Greece

{bgat, ipratika, sper}@iit.demokritos.gr

***Abstract* - In this paper, we propose a novel methodology for text detection in natural scene images. The proposed methodology is based on an efficient binarization and enhancement technique followed by a suitable connected component analysis procedure. Image binarization successfully processes natural scene images having shadows, non-uniform illumination, low contrast and large signal-dependent noise. Connected component analysis is used to define the final binary images that mainly consist of text regions. The proposed methodology results in increased success rates for commercial OCR engines. Experimental results based on a public database of natural scene images prove the efficiency of the proposed approach.**

Index Terms – Text Detection, Scene Text Recognition, Image Binarization, Connected Components Analysis, OCR

I. INTRODUCTION

Natural scene images contain text information which is often required to be automatically recognized and processed. This paper strives toward a novel methodology that aids automatic detection, segmentation and recognition of visual text entities in complex natural scene images. Scene text may be any textual part of the scene images such as street signs, name plates or even text appearing on T-shirts. The research field of scene text recognition receives a growing attention due to the proliferation of digital cameras and the great variety of potential applications, as well. Such applications include robotic vision, image retrieval, intelligent navigation systems and applications to provide assistance to visual impaired persons.

Natural scene images usually suffer from low resolution and low quality, perspective distortion and complex background [1]. Scene text is hard to detect, extract and recognize since it can appear with any slant, tilt, in any lighting, upon any surface and may be partially occluded. Many approaches for text detection from natural scene images have been proposed recently.

Ezaki et al. [2] propose four character extraction methods based on connected components. The performance of the different methods depends on character size. The most effective extraction method proves to be the sequence: Sobel edge detection, Otsu binarization, connected component extraction and rule-based connected component filtering. Yamaguchi et al. [3] propose a digits classification system to recognize telephone numbers written on signboards. Candidate regions of digits are extracted from an image through edge extraction, enhancement and labeling. Since the digits in the images often have skew and slant, the digits are recognized after the skew and slant correction. To correct the skew, Hough transform is used, and the slant is corrected using the method of circumscribing digits with tilted rectangles. In the work of Matsuo et al. [4] a method is proposed that extracts text from scene images after an identification stage of a local target area and adaptive thresholding. Yamaguchi and Maruyama [5] propose a method to extract character regions in natural scene images by hierarchical classifiers. The hierarchy consists of two types of classifiers: a histogram-based classifier and SVM. Finally, Yang et al. [6] have proposed a framework for automatic detection of signs from natural scenes. The framework considers critical challenges in sign extraction and can extract signs robustly under different conditions (image resolution, camera view angle, and lighting).

In this paper, we propose a novel methodology for text detection in natural scene images. The proposed methodology is based on an efficient binarization and enhancement technique followed by a suitable connected component analysis procedure. Image binarization successfully processes natural scene images having shadows, non-uniform illumination, low contrast and large signal-dependent noise. Connected component analysis is used to define the final binary images that mainly consist of text regions. Experimental results show that by using the proposed method we achieve an improved recognition rate for natural scene images.

Our paper is structured as follows: Section 2 is dedicated to a detailed description of the proposed methodology. The experimental results are given in Section 3 while conclusions are drawn in Section 4.

II. METHODOLOGY

The proposed methodology for text detection in natural scene images is based on an efficient binarization and enhancement technique followed by a connected component analysis procedure. The flowchart of the proposed methodology is presented in Figures 1 and 2. Starting from the scene image, we produce gray level image O^1 and inverted gray level image O^{-1}. Then, we calculate the two corresponding binary images I^1 and I^{-1} using an adaptive binarization and image enhancement technique. In the sequel, the proposed technique involves a decision function that indicates which image between binary images I^1 and I^{-1} contains text information. In Fig. 1 the original binary image is selected while in Fig. 2 the inverted binary image is selected. Finally, a procedure that detects connected components of text areas is applied. In the following sections, we describe the main procedures of the proposed methodology.

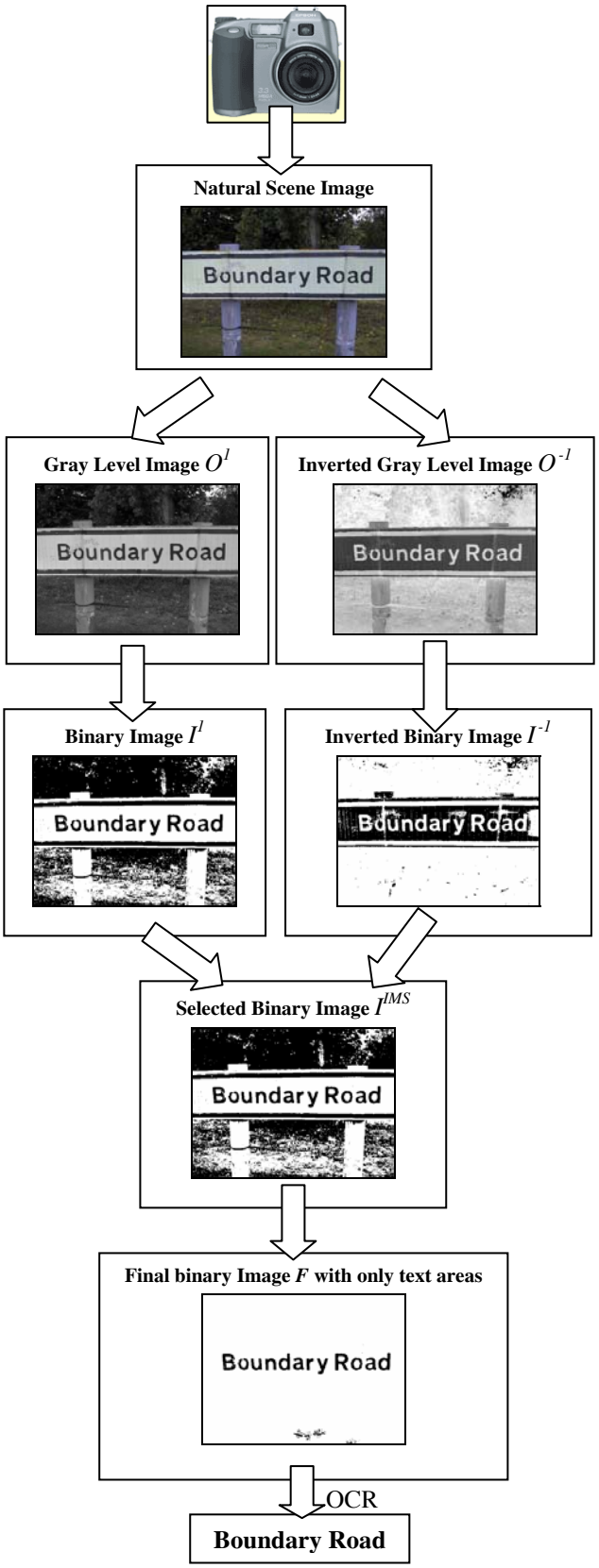


Fig. 1 Flowchart of the proposed method for text detection in natural scene images (*original binary image is selected*).

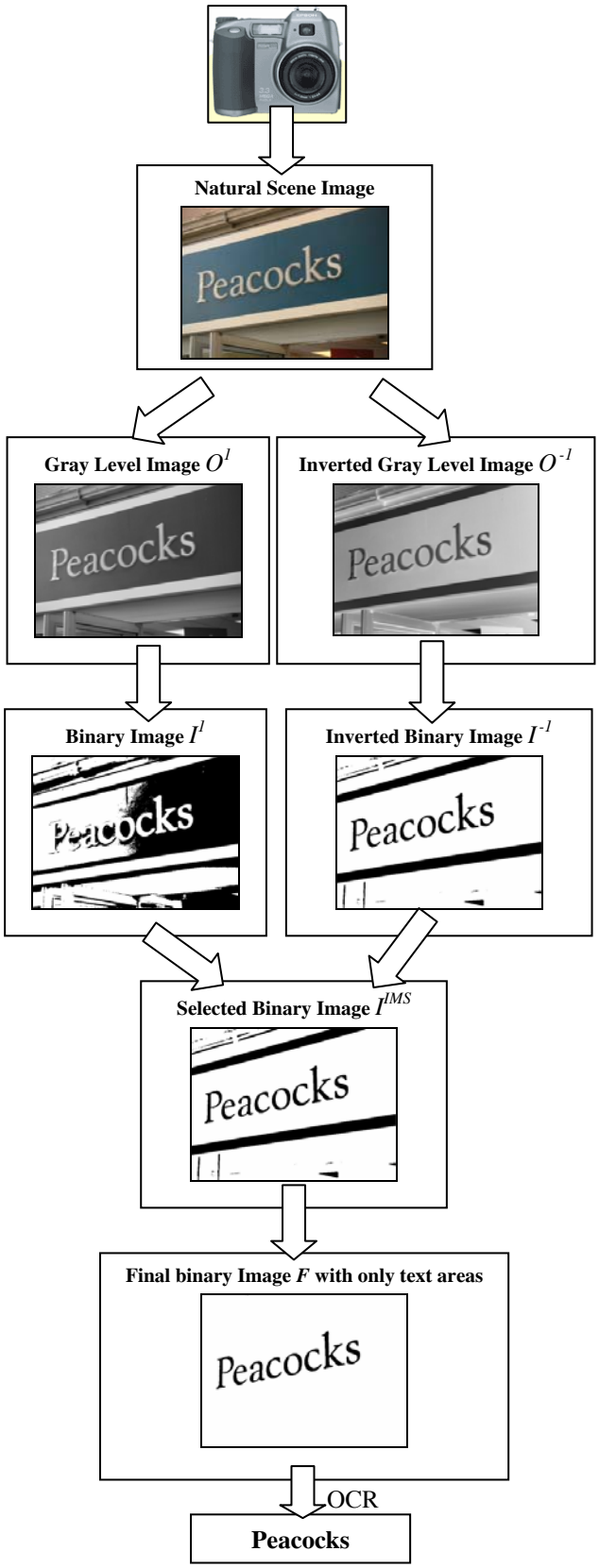


Fig. 2 Flowchart of the proposed method for text detection in natural scene images (*inverted binary image is selected*).

A. *Image Binarization and Enhancement*

Binarization is the starting step of most document image analysis systems and refers to the conversion of the gray-scale or color image to a binary image. Since camera images are most of the times of low quality and low resolution, an image enhancement stage is also essential. The proposed scheme for image binarization and enhancement is mainly based on the work described in [7][8]. It does not require any parameter tuning by the user and can deal with degradations which occur due to shadows, non-uniform illumination, low contrast, large signal-dependent noise, etc. We follow several distinct steps:

1) Image preprocessing: For low resolution and poor quality scene images, a pre-processing stage of the grayscale source image is essential for the elimination of noisy areas, smoothing of background texture as well as contrast enhancement between background and text areas. The use of a low-pass Wiener filter [9] has proved efficient for the above goals. We should mention that we deal with both color and gray scale images. In the case of color images, we use only the luminance component.

2) Rough estimation of foreground regions: At this step, we obtain a rough estimation of foreground regions. Our intention is to proceed to an initial segmentation of foreground and background regions that will provide us with a superset of the correct set of foreground pixels. This is refined at a later step. Sauvola's approach for adaptive thresholding [10] using k = 0.2, is suitable for this case. At this step, we process the original image O(x,y) in order to extract the binary image S(x,y), where 1's correspond to the rough estimated foreground regions.

3) Background surface estimation: At this stage, we compute an approximate background surface $B(x,y)$ of the image O(x,y). Background surface estimation is guided by the valuation of S(x,y) image. For pixels that correspond to 0's in image S(x,y), the corresponding value at B(x,y) equals to O(x,y). For the remaining pixels, the valuation of B(x,y) is computed by a neighboring pixel interpolation. In Fig. 3(b), an example of the estimated background surface of two outdoor scene images is given.

4) Final thresholding: In this step, we proceed to final thresholding by combining the calculated background surface B(x,y) with the original image O(x,y). Text areas are detected if the distance of the preprocessed image O(x,y) from the calculated background B(x,y) exceeds a threshold d. We suggest that the threshold d must change according to the gray-scale value of the background surface B(x,y) in order to preserve textual information even in very dark background areas. For this reason, we use a threshold d that has smaller values for darker regions [8].

5) Image up-sampling: In order to achieve a better quality binary image, we incorporate in the previous step an efficient up-sampling technique. Among available image up-sampling techniques, bi-cubic interpolation is the most common technique that provides satisfactory results [11]. It estimates the value at a pixel in the destination image by an average of 16 pixels surrounding the closest corresponding pixel in the source image.

6) Image post-processing: In the final step, we proceed to post-processing of the resulting binary image in order to eliminate noise, improve the quality of text regions and preserve stroke connectivity by isolated pixel removal and filling of possible breaks, gaps or holes. Our post-processing algorithm involves a successive application of shrink and swell filtering [12].

In Fig. 3(c), an example of the estimated binary images of two outdoor scene images is given. Since scene images may have dark text and light background or vice-versa, it is necessary to process both the original and the negative gray scale image in order to ensure that text information will occur in one of the two resulting binary images.

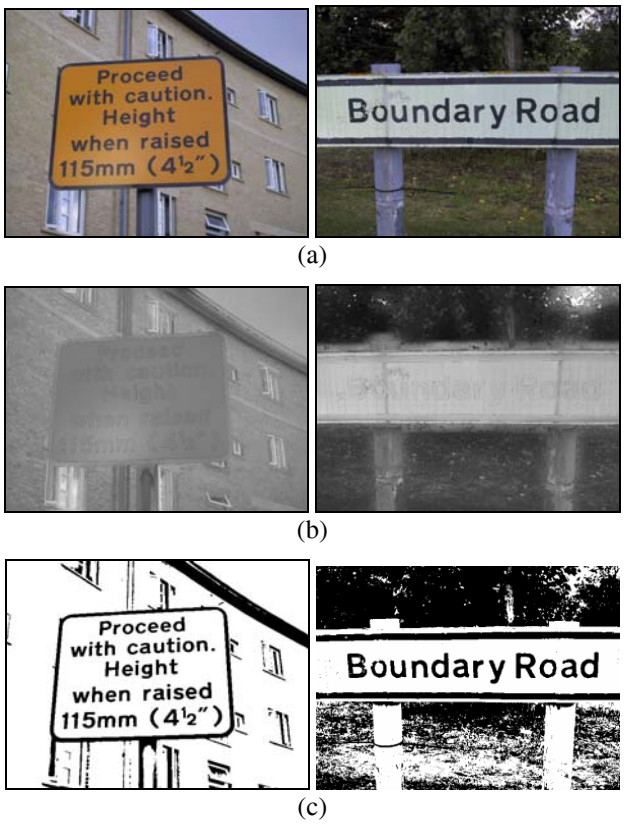


Fig. 3 Image binarization and enhancement example. (a) Original outdoor scene image; (b) Estimated background surface; (c) Resulting image after image binarization and enhancement.

B. Text Areas Detection

After the binarization and enhancement process we get the binary images $I^1(x,y)$ and $I^{-1}(x,y)$, $x \in [1, x_{max}]$, $y \in [1, y_{max}]$. Image I^1 consists of CS^1 connected components C_i^1 with bounding boxes defined by coordinates $[C_{i,x^{TL}}^1, C_{i,y^{TL}}^1]$– $[C_{i,x^{BR}}^1, C_{i,y^{BR}}^1]$, $i \in [1, CS^1]$ while image I^{-1} consists of CS^{-1} connected components C_i^{-1} with bounding boxes defined by coordinates $[C_{i,x^{TL}}^{-1}, C_{i,y^{TL}}^{-1}]$- $[C_{i,x^{BR}}^{-1}, C_{i,y^{BR}}^{-1}]$, $i \in [1, CS^{-1}]$ (see Fig. 4). Function CharOK(C_i^1) determines whether connected component C_i^1 is a character. It takes into account certain limits for the height and width of the connected component along with the appearance of neighboring connected components with almost the same height in the horizontal direction. Function $CharOK(C_i^1)$ is defined as follows:

$$CharOK(C_i^1) = \begin{cases} 1, \text{ if } T1(C_i^1) \textbf{ AND} \\ \quad (\exists j: T2(C_i^1, C_j^1) \textbf{ AND } T3(C_i^1, C_j^1) \\ \quad\quad \textbf{AND } T4(C_i^1, C_j^1)) \\ 0, \text{ otherwise} \end{cases} \quad (1)$$

where

$$T1(C_i^1) = \begin{cases} \text{TRUE, if } (\frac{x_{max}}{MaxChars} < C_{i,x^{BR}}^1 - C_{i,x^{TL}}^1 < \frac{x_{max}}{MinChars}) \\ \quad \textbf{AND } (\frac{y_{max}}{MaxLines} < C_{i,y^{BR}}^1 - C_{i,y^{TL}}^1 < \frac{y_{max}}{MinLines}) \\ \text{FALSE, otherwise} \end{cases} \quad (2)$$

$$T2(C_i^1, C_j^1) = \begin{cases} \text{TRUE, if } (\left| C_{j,x^{TL}}^1 - C_{i,x^{BR}}^1 \right| < 2(C_{i,y^{BR}}^1 - C_{i,y^{TL}}^1) \\ \quad \textbf{OR } \left| C_{j,x^{BR}}^1 - C_{i,x^{TL}}^1 \right| < 2(C_{i,y^{BR}}^1 - C_{i,y^{TL}}^1) \\ \text{FALSE, otherwise} \end{cases} \quad (3)$$

$$T3(C_i^1, C_j^1) = \begin{cases} \text{TRUE, if } \left| C_{j,y^{TL}}^1 - C_{i,y^{TL}}^1 \right| < C_{i,y^{BR}}^1 - C_{i,y^{TL}}^1 \\ \text{FALSE, otherwise} \end{cases} \quad (4)$$

$$T4(C_i^1, C_j^1) = \begin{cases} \text{TRUE, if } \frac{\left| (C_{i,y^{BR}}^1 - C_{i,y^{TL}}^1) - (C_{j,y^{BR}}^1 - C_{j,y^{TL}}^1) \right|}{C_{i,y^{BR}}^1 - C_{i,y^{TL}}^1} < 0.3 \\ \text{FALSE, otherwise} \end{cases} \quad (5)$$

where parameters *MaxChars* and *MinChars* correspond to the maximum and minimum number of expected characters in a text line, *MaxLines* and *MinLines* correspond to the maximum and minimum number of expected text lines and *e* is a small float. In our experiments, we used *MaxChars* = 100, *MinChars* = 5, *MaxLines* = 50, *MinLines* = 3.

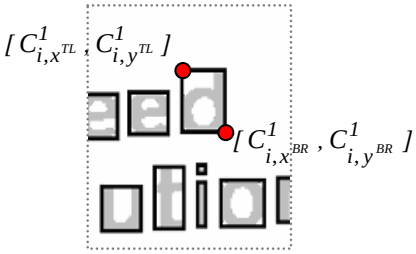


Fig. 4 Connected components of the binary image.

$T1(C_i^1)$ is TRUE if the height and width of the connected component C_i^1 are in between certain limits, $T2(C_i^1, C_j^1)$ is TRUE if the connected components C_i^1 and C_j^1 are neighbors in horizontal direction, $T3(C_i^1, C_j^1)$ is TRUE if the connected components C_i^1 and C_j^1 belong to the same text line, and $T4(C_i^1, C_j^1)$ is TRUE if the connected components C_i^1 and C_j^1 have similar height.

Between the two images I^1 and I^{-1}, we select the one that has more connected components determined as characters. IMS denotes the selected image according to the formula:

$$IMS = \begin{cases} 1, \text{if } \sum_{i=1}^{CS^1} CharOK(C_i^1) > \sum_{i=1}^{CS^{-1}} CharOK(C_i^{-1}) \\ -1, \text{otherwise} \end{cases} \quad (6)$$

The final binary image F consists of all the connected components of image I^{IMS} that are detected as characters, that is $CharOK(C_i^{IMS}) = 1$, as well as their adjacent ones. This is done, in order to include broken characters and character parts of small height. Binary image F is given by the following formula:

$$\begin{aligned} F = \cup C_i^{IMS}, i: \; & CharOK(C_i^{IMS}) = 1 \textbf{ OR} \\ & \exists j: (CharOK(C_j^{IMS}) = 1 \textbf{ AND} \\ & \sqrt{(C_{j,x^{TL}}^{IMS} - C_{i,x^{TL}}^{IMS})^2 - (C_{j,y^{TL}}^{IMS} - C_{i,y^{TL}}^{IMS})^2} + \\ & \sqrt{(C_{j,x^{BR}}^{IMS} - C_{i,x^{BR}}^{IMS})^2 - (C_{j,y^{BR}}^{IMS} - C_{i,y^{BR}}^{IMS})^2} < 4(C_{i,y^{BR}}^{IMS} - C_{i,y^{TL}}^{IMS})) \end{aligned} \quad (7)$$

In Fig. 5, a text detection example is demonstrated. In Fig. 5(b) and Fig. 5(c) the original and the inverted binary images as well as their connected components are shown, respectively. Between these two images we select image of Fig. 5(c) due to criterion in Eq. 6. The final binary image is shown in Fig. 5(d) and consists of the detected connected components as characters in the selected binary image.

III. Experimental Results

The proposed algorithm was tested using the public database of the ICDAR2003 Robust Reading Competition [14]. We focused on the dataset for "Robust Reading and Text Locating" competition. A list of some representative results is presented in Table 1. In almost all the cases, the

text areas are detected in the final binary images while the non-text areas are eliminated. The proposed method worked successfully even in cases with degradations, shadows, non-uniform illumination, low contrast and large signal-dependent noise. An experiment to automatically quantify the efficiency of the proposed text detection method was also performed. We compared the results obtained by the well-known OCR engine ABBYY FineReader_6 [15] with and without the incorporation of the proposed technique. To quantify the OCR results we calculated the Levenshtein distance [16] between the correct text (ground truth) and the resulting text for several scene images. As shown in the representative results at Table 2, the application of the proposed text detection technique has shown best performance with respect to the final OCR results. Total results show that by using the proposed method we achieve a more than 50% improvement of the FineReader_6 recognition rate for natural scene images.

IV. Conclusion and Future Work

This paper strives toward a novel methodology that aids automatic detection, segmentation and recognition of visual text entities in complex natural scene images. The proposed methodology is based on an efficient binarization and enhancement technique followed by a suitable connected component analysis procedure. Image binarization successfully processes natural scene images having shadows, non-uniform illumination, low contrast and large signal-dependent noise. Connected component analysis is used to define the final binary images that mainly consist of text regions. Experimental results show that by using the proposed method we achieve an improved recognition rate for natural scene images.

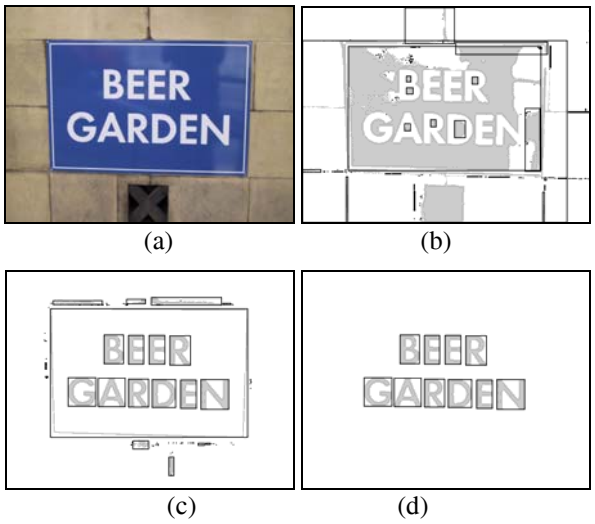


Fig. 5 An example of text area detection. (a) Original scene image; (b) Resulting binary image I^1; (c) Resulting inverted binary image I^{-1}; (d) Resulting final binary image F with the detected text areas. In all cases, the surrounding boxes of the connected components are shown.

TABLE I
Text detection representative results

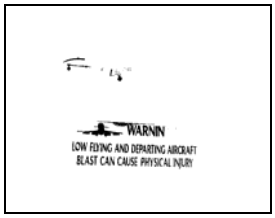
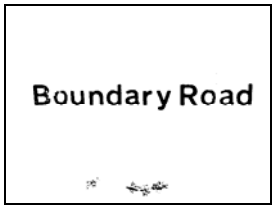
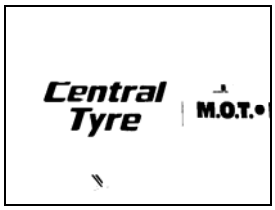
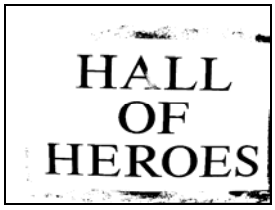
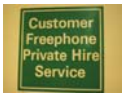
Natural scene image	Text detection result
	
	
	
	
	
	
	

TABLE II
OCR RESULTS WITH AND WITHOUT THE INCORPORATION OF THE PROPOSED METHOD

Natural scene image	Levenshtein Distance from the Ground truth		Natural scene image	Levenshtein Distance from the Ground truth	
	FineReade6	Proposed method+FineReade6		FineReade6	Proposed method+FineReade6
	21	0		1	1
	25	18		2	2
	5	4		32	3
	2	2		2	0
	3	3		39	18
	1	1		10	1
	0	0		10	10
	4	1		6	3
	0	0		38	16
			TOTAL	**201**	**83**

REFERENCES

[1] David Doermann, Jian Liang, Huiping Li, "Progress in Camera-Based Document Image Analysis", In Proceedings of Seventh International Conference on Document Analysis and Recognition (ICDAR'03), 2003, pp. 606-616.

[2] N. Ezaki, M. Bulacu and L. Schomaker, "Text detection from natural scene images:Towards a system for visually impaired persons", In Proceedings of the International Conference on Pattern Recognition (ICPR'04), 2004, pp. 683-686.

[3] 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, Edinburgh, Scotland, 3-6 August 2003, pp. 359–363.

[4] 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), February 2002, pp. 232–241.

[5] T. Yamaguchi and M. Maruyama, "Character Extraction from Natural Scene Images by Hierarchical Classifiers", In Proceedings of the International Conference on Pattern Recognition (ICPR'04), 2004, pp. 687-690.

[6] 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, pp, 1-8.

[7] B. Gatos, I. Pratikakis and S.J. Perantonis, "Locating text in historical collection manuscripts", Lecture Notes on AI, SETN, 2004, pp 476-485.

[8] B. Gatos, I. Pratikakis and S.J. Perantonis, "An adaptive binarisation technique for low quality historical documents", IARP Workshop on Document Analysis Systems (DAS2004), Lecture Notes in Computer Science (3163), September 2004, pp. 102-113.

[9] A. Jain, Fundamentals of Digital Image Processing, Prentice Hall, Englewood Cliffs, NJ (1989).

[10] J. Sauvola, M. Pietikainen, Adaptive Document Image Binarization, Pattern Recognition 33, 2000, pp. 225-236.

[11] H. Hsieh, H. Andrews, "Cubic splines for image interpolation and digital filtering", IEEE Transactions on Acoustics, Speech, and Signal Processing, 26(6), 1978, pp. 508-517.

[12] R.J. Schilling, *Fundamentals of Robotics Analysis and Control*, Prentice-Hall, Englewood Cliffs, NJ (1990)

[13] Robust Reading Competition Database, http://algoval.essex.ac.uk/icdar/Datasets.html, 2005.

[14] ABBYY, www.finereader.com, 2005.

[15] V.I. Levenshtein, "Binary codes capable of correcting deletions, insertions and reversals", Sov. Phys. Dokl., 6. 1966, pp. 707-710.