Improving Computer Vision-Based Indoor Wayfinding for Blind Persons with Context Information

YingLi Tian[1], Chucai Yi[1], and Aries Arditi[2]

[1] Electrical Engineering Department
The City College and Graduate Center
City University of New York, New York, NY 10031
ytian@ccny.cuny.edu, cyi@gc.cuny.edu
[2] Arlene R Gordon Research Institute
Lighthouse International
111 East 59th Street, New York, NY 10022
aarditi@lighthouse.org

Abstract. There are more than 161 million visually impaired people in the world today, of which 37 million are blind. Camera-based computer vision systems have the potential to assist blind persons to independently access unfamiliar buildings. Signs with text play a very important role in identification of bathrooms, exits, office doors, and elevators. In this paper, we present an effective and robust method of text extraction and recognition to improve computer vision-based indoor wayfinding. First, we extract regions containing text information from indoor signage with multiple colors and complex background and then identify text characters in the extracted regions by using the features of size, aspect ratio and nested edge boundaries. Based on the consistence of distances between two neighboring characters in a text string, the identified text characters have been normalized before they are recognized by using off-the-shelf optical character recognition (OCR) software products and output as speech for blind users.

Keywords: Indoor navigation and wayfinding, indoor, computer vision, text extraction, optical character recognition (OCR).

1 Introduction

Based on the 2002 world population, there are more than 161 million visually impaired people in the world today, of which 37 million are blind [1]. Challenges associated with independent mobility are well known to reduce quality of life and compromise the safety of individuals with severe vision impairment. Robust and efficient indoor object detection from images/video captured by a wearable camera has the potential to help people with severe vision impairment to independently access unfamiliar indoor environments. But indoor object detection poses several challenges: 1) there are large intra-class variations of the object model among different indoor environments. The appearance and design style of different instances of objects (e.g. doors) are quite variable in different buildings; 2) there are small

K. Miesenberger et al. (Eds.): ICCHP 2010, Part II, LNCS 6180, pp. 255–262, 2010.

inter-class variations for certain object models. As shown in Figure 1, the basic door shapes of a bathroom, an exit, a lab door, and an elevator are very similar. It is very difficult for a machine-based system to distinguish them without using context information on/around the object; 3) in contrast to objects with enriched texture and color in natural or outdoor scenes, most indoor objects are man-made with less texture. Existing feature descriptors may not effectively represent the indoor objects; 4) it is unrealistic to expect an intact object, especially in its entirety and in a canonical orientation, to be captured by a blind user with a wearable camera that has no reliable means for steering or aiming the camera. Indoor object detection methods for our application must handle situations in which only part of the object is captured; and 5) when the user moves while wearing a camera, changes of position and distance between the user and the object will cause big view and scale variations of the objects.

Signs with text and icons play an important role in guiding us through unknown environments and can be very helpful to persons who are blind, if the signage can be recognized and decoded by a wearable computer vision system. To improve the ability of people who are blind or have significant visual impairments to access, understand, and explore unfamiliar indoor environments, it is essential to incorporate the context information with object detection from multiple visual cues captured using wearable cameras. In this paper, we present a robust and efficient algorithm for text extraction from indoor signage. The extracted text is recognized by the off-the-shelf optical character recognition (OCR) software products and then converted to speech for blind users by existing speech-synthesis software.

Fig. 1. Indoor objects (top row) and their associated contextual information (bottom row): (a) a bathroom, (b) an exit, (c) a lab room, (d) an elevator. Contextual information (bottom row) must be employed to distinguish them.

2 State-of-the-Art

Camera-based document analysis and recognition have recently gained more attention due to the pervasive use of camera phones and hand-held digital still and video cameras [2-5, 7-8, 14]. Compared to high-resolution and high-quality scanned document images, camera-captured document images will result in poor performance due to: (1) Insufficient resolution, (2) uneven lighting, (3) perspective and camera lens distortion, (4) text on non-planar surfaces, (5) complex backgrounds, (6) defocus (8) camera or document movement and image smear, (9) intensity and color quantization, and (10)

sensor noise. Liang et al. described these challenges with more details in a well-known survey paper [4]. Many methods have been proposed for text extraction, enhancement, and binarization. A survey conducted by Trier and Taxt compared 11 locally adaptive thresholding techniques [10]. They concluded that Niblack's method [6] is the most effective for general images. However, the method produces a noisy output with homogeneous regions larger than the size of the window. Some improved methods [9, 12] have been proposed by assuming black text on a white background or by measuring local contrast to estimate threshold. However, these conventional methods cannot handle text on cluttered backgrounds. Recently, a number of papers have addressed text extraction from colorful documents [3, 7, 15]. Kasar et al. [3] proposed a straightforward method to robustly extract the text from cluttered backgrounds and convert the text in black on a white background. Nikolaou et al. [7] developed a meanshift method to reduce the number of complex background colors for document analysis.

3 Context Information Extraction from Signage

Signage and other visual information provide important guidance in finding a route through a building by showing the direction and location of the desired end-point. Visual information is particularly important for distinguishing similarly shaped objects such as elevators, bathrooms, exits, and normal doors. Figure 2 shows some examples of signage of indoor environments. In this paper, we focus on text extraction and recognition from signage. Our method includes three components: 1) color reduction; 2) text extraction and normalization; and 3) text recognition.

Fig. 2. Examples of signage in indoor environments

3.1 Color Reduction for Text Extraction

To segment and extract text from images containing multiple colors, inspired by [7], we develop a color reduction method as a quantization preprocessing stage based on edge and color information. First, an edge preserving smoothing filter is applied to the entire image to remove image noises without deforming the details of object boundaries. Second, in order to represent the colors of all the objects on the image, the RGB color distribution of the image is sampled by selecting only those pixels which are local minima in the 8-neighborhood on the edge map image. This process ensures that the samples are non-edge points, which guarantees that they are not on the boundary

of the objects. Thus, fuzzy points on the transition areas between objects on the image are avoided. Third, the sampled pixels with similar colors are grouped into color clusters. Fourth, a mean-shift operation is performed on the obtained color clusters to further merge similar color regions in RGB space and produce the final number of colors. This final number of colors should be very small (< 10). At this point, the processed image should have solid characters and uniform local backgrounds for text extraction.

After color quantization, with the purpose of classifying the information with different colors, each quantized value is placed against a white background, as shown in Figure 3. Obviously color reduction decomposes image into several color layers. Each layer contains exactly two colors in the form of informative foreground and white background. The interference from background is reduced for text extraction.

Fig. 3. Examples of color reduction. The 1st column shows the original images; the 2nd and 3rd columns show color layers after color quantization.

3.2 Text Extraction

To extract the regions containing text from cluttered backgrounds, we apply a text extraction method in each color layer which is obtained from color reduction processing. Generally, text regions have more transitions between foreground (e.g. text) and background than non-text regions. Based on this observation, the maximum gradient difference (MGD) [13] for each pixel is calculated to describe the density of binary switches, which are transitions from foreground to background and vice versa in each color layer of the input image, where the foreground is denoted by 1 and the background is denoted by 0. A *1×9* sliding window is defined to detect the transitions around each pixel, as shown in Figure 4(a). The MGD of each pixel is calculated by Equation (1).

$$MGD(P_c) = \begin{cases} 0, & \sum_{k=-4}^{4} P_{c+k} = 0 \ or \ \prod_{k=-4}^{4} P_{c+k} = 1 \\ 1, & otherwise \end{cases} \quad (1)$$

where P_c is a pixel located in the center of the sliding window $\{P_{c+k} | -4 \le k \le 4\}$.

In the MGD map, regions containing text information correspond to clusters of value 1. Therefore we search the MGD map for the rows and columns which compose text regions by summing the MGD values row-by-row and column-by-column, as presented by Equation (2).

$$R_H(i) = \sum_j MGD\left(P_c(i,j)\right), \quad R_V(j) = \sum_{R_H(i)>T_H} MGD\left(P_c(i,j)\right) \tag{2}$$

where $R_H(i)$ is the accumulation of MGD values in the *ith* row and $R_V(j)$ is that in the *jth* column. If $R_H(i)$ and $R_V(j)$ are greater than predefined thresholds T_H and T_V respectively, the corresponding rows and columns are components of text regions, as shown in Figure 4(b).

In order to detect more complex text with irregular size and shape, we extend the algorithm by using a text structure model to identify letters and numbers. This model is based on the fact that each of these characters will contain no more than two closed edge boundaries (e.g., "A" contains one such boundary and "8" contains two). First, the edge map of the text regions is computed, as shown in Figure 5(c-d). The edge bounding box is obtained by performing connected component labeling. A simple ratio filter is then applied to filter out the obvious non-text regions, as shown in Figure 5(e-f). To convert the text to black on a white background, Kasar *et al.* [3] employed an edge-based connected component approach to automatically determine a threshold for each component of RGB channels. In our experiment, each color layer has already been a binary image, so we only set the foreground pixels as black and background as white.

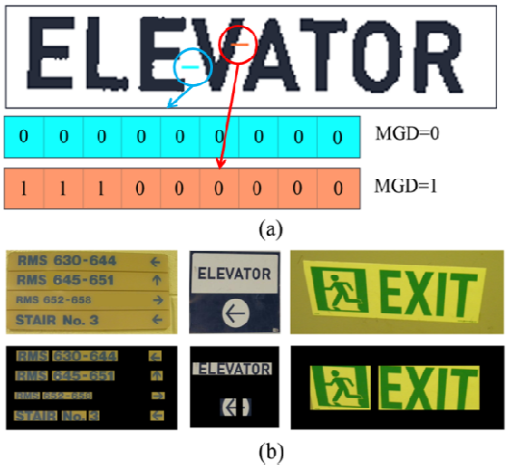

Fig. 4. (a) Examples of sliding windows in a text region; (b) Text regions detected by MGD algorithm. Top row is original image; bottom row shows identified text regions against a black background.

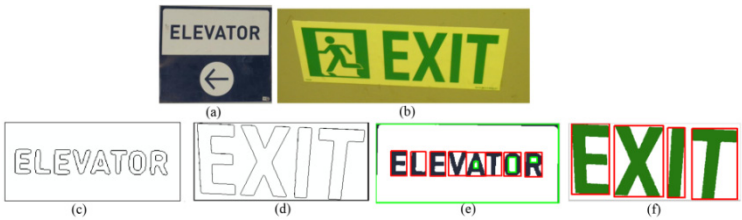

Fig. 5. (a-b) Original images; (c-d) Edge map of some text regions; (e-f) Edge bounding boxes where the red boxes indicate the regions of text characters. The green boxes are the non-text regions which filtered out by predefined rules.

3.3 Text Normalization and Recognition

Off-the-shelf optical character recognition (OCR) software for text recognition has constraints with respect to font size and orientation variability. Before we input the text image to OCR, we must normalize the text to the size range which these products can handle. We use more text structure features based on the geometric relationships among three neighboring characters to further filter out the non-text bounding boxes and normalize text characters. As shown in Figure 6, the geometric relationship includes the consistency of distances between centroids. The normalized text strings are then recognized by OCR and displayed as speech to blind users.

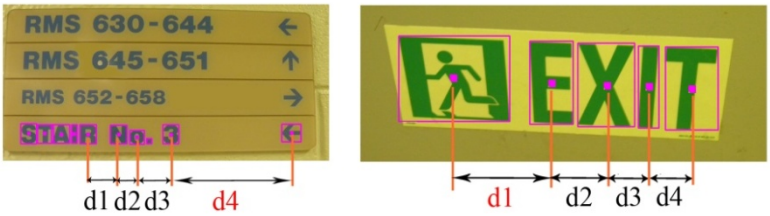

Fig. 6. Text bounding boxes share consistent distances while non-text bounding boxes are connected by inconsistent distance denoted by red

4 Experiment Results

Our method is robust and efficient for text extraction and recognition and is evaluated on a database of signs in indoor environments. Figure 7 shows some examples of the results for text extraction and recognition from indoor signage. The extracted text characters are recognized by using an off-the-shelf OCR software product. In our experiment, OmniPage Professional Version 16 is used to transform the extracted text from image regions to readable ASCII codes.

Table 1 demonstrates the results of OCR detection corresponding to Figure 7. The context information, including both letters and digits, is accurately extracted by our method. Note that our text extraction makes using standard OCR feasible. Images of signs from a camera without this process would surely fail or provide inaccurate character recognitions.

Table 1. Results of text recognition of images in Figure 7.

Text images	OCR outputs
Figure 7(a)	RMS 630-644 RMS 645-651 RMS 652-658 STAIR NO.3
Figure 7(b)	EXIT
Figure 7(c)	WOMEN
Figure 7(d)	LABORATORY
Figure 7(e)	ELEVATOR

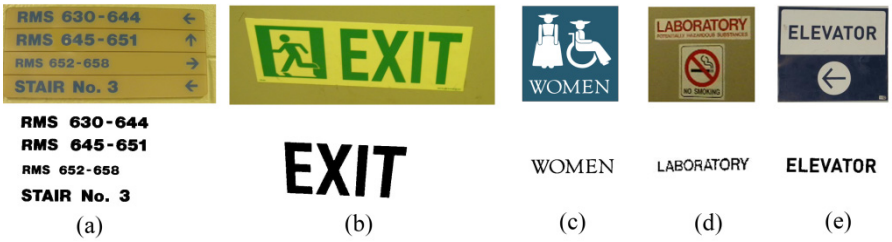

Fig. 7. Examples results of text extraction for signage in indoor environments

5 Impact and Contributions

We have developed a new framework for text extraction and recognition. Context information, including that provided by indoor signage, plays a very important role for navigation and wayfinding of blind or visually impaired people independently accessing unfamiliar indoor environments. This research has following impact: (1) It significantly enriches the study of object detection by incorporating context information, and leads to significant improvements over existing methods; (2) The method developed in this paper provides new strategies and the promise of new technologies for blind and visually impaired persons to access unfamiliar indoor environments; and (3) The research will benefit many other important research areas including video surveillance, intelligent conference rooms, video indexing, human-computer interactions, *etc.*

6 Conclusion and Future Work

We have described a new method to extract and recognize text from indoor signage. First, we propose a color reduction process to reduce complex colors of both text and background. Then, we apply a robust and efficient text extraction to generate black text on white background. Finally, to achieve high recognition by using existing OCR software, we apply a post-processing step to model geometric relationship among neighbor characters to further filter out the non-text bounding boxes and use the consistency of distances between centroids of these text characters. Our future work will focus on recognizing iconic signage, which of course does not contain text

information. We will also address the significant human interface issues associated with prioritizing which objects and signs are most relevant to the blind user's navigation goals, and with the significant issues relating to auditory display of the recognized information.

Acknowledgments. This work was supported by NSF grant IIS-0957016 and NIH grant EY017583.

References

1. Resnikoff, S., Pascolini, D., Etya'ale, D., Kocur, I., Pararajasegaram, R., Pokharel, G.P., et al.: Global data on visual impairment in the year 2002. Bulletin of the World Health Organization 82, 844–851 (2004)
2. Fu, H., Liu, X., Jia, Y.: Gaussian Mixture Modeling of Neighbor Characters for Multilingual Text Extraction in Images. In: IEEE International Conference on Image Processing, ICIP (2006)
3. Kasar, T., Kumar, J., Ramakrishnan, A.G.: Font and Background Color Independent Text Binarization. In: 2nd International Workshop on Camera-Based Document Analysis and Recognition (2007)
4. Liang, J., Doermann, D., Li, H.: Camera-based analysis of text and documents: a survey. IJDAR 7(2-3) (July 2005)
5. Lyu, M., Song, J., Cai, M.: A Comprehensive method for multilingual video text detection, localization, and extraction. IEEE transactions on circuits and systems for video technology 15 (2005)
6. Niblack, W.: An introduction to digital image processing, pp. 115–116. Prentice Hall, Englewood Cliffs (1986)
7. Nikolaou, N., Papamarkos, N.: Color Reduction for Complex Document Images. International Journal of Imaging Systems and Technology 19, 14–26 (2009)
8. Phan, T., Shivakumara, P., Lim Tan, C.: A Laplacian Method for Video Text Detection. In: The 10th International Conference on Document Analysis and Recognition, pp. 66–70 (2009)
9. Sauvola, J., Pietikainen, M.: Adaptive document image binarization. Pattern Recognition 33, 225–236 (2000)
10. Trier, O., Taxt, T.: Evaluation of binarization methods for document images. IEEE Trans. Pattern Analysis and Machine Intelligence 17(3), 312–315 (1995)
11. Tran, H., lux, A., Nguyen, H., Boucher, A.: A novel approach for text detection in images using structural features. In: The 3rd International Conference on Advances in Pattern Recognition, pp. 627–635 (2005)
12. Wolf, C., Jolion, J., Chassaing, F.: Text localization, enhancement and binarization in multimedia documents. In: Proc. ICPR, vol. 4, pp. 1037–1040 (2002)
13. Wong, E., Chen, M.: A new robust algorithm for video text extraction PR36(6) (June 2003)
14. Zhang, J., Kasturi, R.: Extraction of Text Objects in Video Documents: Recent Progress. In: The 8th IAPR International Workshop on Document Analysis Systems (2008)
15. Zhong, Y., Karu, K., Jain, A.K.: Locating text in complex color images. In: Proc. ICDAR (1995)