Context-based Indoor Object Detection as an Aid to Blind Persons Accessing Unfamiliar Environments

Xiaodong Yang, Yingli Tian
Dept. of Electrical Engineering
The City College of New York
NY 10031, USA
{xyang02, ytian}@ccny.cuny.edu

Chucai Yi
The Graduate Center
The City University of New York
NY 10016, USA
cyi@gc.cuny.edu

Aries Arditi
Arlene R Gordon Research Institute
Lighthouse International
New York, NY 10222, USA
aarditi@lighthouse.org

ABSTRACT
Independent travel is a well known challenge for blind or visually impaired persons. In this paper, we propose a computer vision-based indoor wayfinding system for assisting blind people to independently access unfamiliar buildings. In order to find different rooms (i.e. an office, a lab, or a bathroom) and other building amenities (i.e. an exit or an elevator), we incorporate door detection with text recognition. First we develop a robust and efficient algorithm to detect doors and elevators based on general geometric shape, by combining edges and corners. The algorithm is generic enough to handle large intra-class variations of the object model among different indoor environments, as well as small inter-class differences between different objects such as doors and elevators. Next, to distinguish an office door from a bathroom door, we extract and recognize the text information associated with the detected objects. We first extract text regions from indoor signs with multiple colors. Then text character localization and layout analysis of text strings are applied to filter out background interference. The extracted text is recognized by using off-the-shelf optical character recognition (OCR) software products. The object type, orientation, and location can be displayed as speech for blind travelers.

Categories and Subject Descriptors
I.4.8 [**Scene Analysis**]: Object Recognition

General Terms
Algorithms, Design.

Keywords
Indoor wayfinding, computer vision, object detection, text extraction, blind/visually impaired persons.

1. INTRODUCTION
Robust and efficient indoor object detection can help people with severe vision impairment to independently access unfamiliar indoor environments. While GPS-guided electronic wayfinding aids show much promise in outdoor environments, there is still a lack of orientation and navigation aids to help people with severe vision impairment to independently find doors, rooms, elevators, stairs, bathrooms, and other building amenities in unfamiliar indoor environments. Computer vision technology has the potential to assist blind individuals to independently access, understand, and explore such environments.

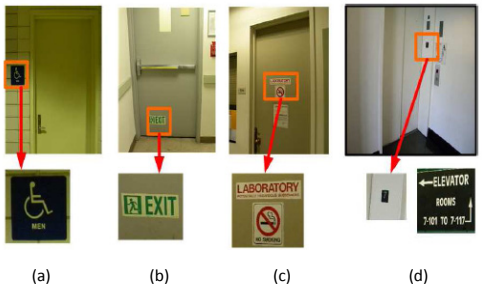

Figure 1: Typical indoor objects (top row) and their associated contextual information (bottom row). (a) a bathroom, (b) an exit, (c) a laboratory, (4) an elevator.

Computer vision-based indoor object detection is a challenging research area due to following factors: 1) there are large intra-class variations of appearance and design of objects in different architectural environments; 2) there are relatively small inter-class variations of different object models. As shown in Fig. 1, the basic shapes of a bathroom, an exit, a laboratory, and an elevator are very similar. It is very difficult to distinguish them without using the associated context information; 3) compared to objects with enriched texture and color in natural scene or outdoor environments, most indoor objects are man-made with less texture. Existing feature descriptors which work well for outdoor environments may not effectively represent indoor objects; and 4) when a blind user moves with wearable cameras, the changes of position and distance between the user and the object will cause big view variations of the objects, as well as only part of the objects is captured. Indoor wayfinding aid should be able to handle object occlusion and view variations.

To improve the ability of people who are blind or have significant visual impairments to independently access, understand, and explore unfamiliar indoor environments, we propose a new framework using a single camera to detect and recognize doors, and elevators, incorporating text information associated with the detected object. In order to discriminate similar objects in indoor environments, the text information associated with the detected objects is extracted. The extracted text is then recognized by using off-the-shelf optical character recognition (OCR) software. The object type and relative position are displayed as speech for blind travelers.

MM'10, October 25–29, 2010, Firenze, Italy.

2. INDOOR OBJECT DETECTION

2.1 Door Detection

Among indoor objects, doors play a significant role for navigation and provide transition points between separated spaces as well as entrance and exit information. Most existing door detection approaches for robot navigation employed laser range finders or sonar to obtain distance data to refine the detection results from cameras [1, 3, 9]. However, the problem of portability, high-power, and high-cost in these systems with complex and multiple sensors make them inappropriate to work for visually impaired people. To reduce the cost and complexity of the device and computation and enhance the probability, we use a single camera. A few algorithms using monocular visual information have been developed for door detection [2, 7]. Most existing algorithms are designed for specific environments, which cannot be extended to unfamiliar environments. Some cannot discriminate doors from typical large rectangular objects in indoor environments, such as bookshelves, cabinets, and cupboards.

In this paper, our door detection method is extended based on the algorithm in our paper [11]. We built a general geometric model which consists of four corners and four bars. The model is robust to variations in color, texture, and occlusion. Considering the diversity and variance of appearance of doors in different environments, we utilized the general and stable features, edges and corners to implement door detection.

2.2 Relative Position Determination

For indoor wayfinding device to assist blind or visually impaired persons, we need also, after detecting the door, to indicate the door's position to the blind user. We classify the relative position as *Left*, *Front*, and *Right* as shown in Fig. 2. *LL* and *LR* correspond to the length of left vertical bar and the length of right vertical bar of a door in an image. From the perspective projection, if a door locates on the left side of a camera, then *LL* is larger than *LR*; if a door locates on the right side of a camera, then *LL* is smaller than *LR*; if a door locates in front of a camera, the difference between *LL* and *LR* is smaller than a threshold. Therefore, we can infer the relative position of a door with respect to the observer by comparing the values of *LL* and *LR*. In practice, to avoid scale variance, we use the angle formed by the horizontal bar and the horizontal axis of an image to determine the relative position of a door.

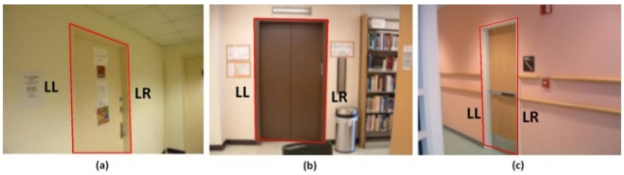

Figure 2: Examples of door position determination: (a) *Left*, (b) *Front*, and (c) *Right*.

2.3 Distinguish Concave and Convex Objects

The depth information plays an important role in indoor objects recognition, as well as in robotics, scene understanding and 3-D reconstruction. In most cases, doors are receded into a wall, especially for the doors of elevators. Other large objects with door-like shape and size, such as bookshelves and cabinets, extend outward from a wall. In order to distinguish concave (e.g. doors and elevators) and convex (e.g. bookshelves, cabinets) objects, we propose a novel and simple algorithm by integrating the lateral information of a detected doorframe to obtain the depth information with respect to the wall.

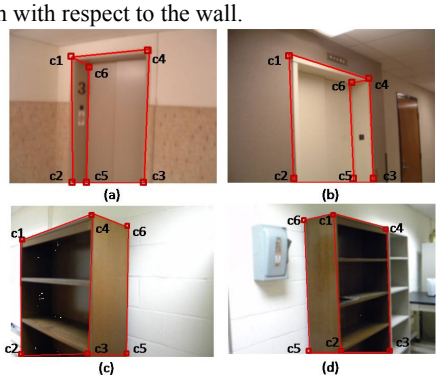

Figure 3: Concave and convex models. (a-b) concave object; (c-d) convex objects.

Due to the different geometric characteristics and different relative positions, concave and convex objects demonstrate the laterals in different positions with respect to the detected "doorframe", as shown in Fig. 3. In Fig. 3(a), since *LL*<*LR*, the elevator door locates on the right side with respect to the user. Furthermore, the elevator door, a concave object, demonstrates its lateral (C_1-C_2-C_5-C_6) on the left side of the doorframe (C_1-C_2-C_3-C_4). In Fig. 3(c), since *LL*<*LR*, the bookshelf locates on the right side and as a convex object, its lateral (C_4-C_3-C_5-C_6) on the right side of the frame (C_1-C_2-C_3-C_4). Similar relations can be found in Fig. 3(b) and (d). Therefore, combining the position of a "doorframe" and the position of its lateral, we can determine the convexity or concavity of the "doorframe", as shown in Table 1. Note that the position of a frame is relative to the user; the position of a lateral is relative to the frame.

Table 1. Convexity or concavity determination combining the position of a frame and the position of a lateral.

Position of a frame	Position of a lateral	Convexity or Concavity
Left	Left	Convex
Left	Right	Concave
Right	Left	Concave
Right	Right	Convex

3. ALGORITHM OF TEXT EXTRACTION, LOCALIZATION, AND RECOGNITION

The context information includes different kinds of visual features. Our system focuses on finding and extracting text from signage. Many efforts have been made to extract text regions from the scene image [4, 6, 8, 10]. We propose a robust algorithm to extract text from complex background with multiple colors. Structural features of single text character and layout analysis of text strings containing at least 3 character members are both employed to refine the detection and localization of text region.

3.1 Text Region Extraction

3.1.1 Color Decomposition

In order to extract text information from indoor signage with complex background, we first design a robust algorithm to extract the coarse regions of interest (ROI) that are very likely to contain

text information. We observe that each text string generally has a uniform color in most indoor environments. Thus color decomposition is performed to obtain image regions with identical colors to different color layers. Each color layer contains only one foreground color with a white background, as shown in Fig. 4.

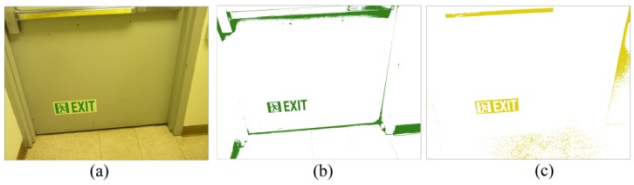

Figure 4: Two color layers are extracted from an exit door. (a) the original image; (b-c) different color layers extracted by color decomposition.

3.1.2 Text Region Extraction

After color decomposition, we extract text regions based on the structure of text characters. Text characters which are composed of strokes and major arcs are well-aligned along direction of the text string they belong to. In each color layer, the well structured characters result in high frequency of binary switches from foreground to background and vice versa. Generally, text strings on indoor signage have layouts with a dominant horizontal direction respect to the image frame. We apply a $1 \times N$ sliding window to calculate the binary switch around the center pixel of the window P. In our system, we choose $N = 13$. If there are no less than 2 binary switches occurring inside the sliding window, the corresponding central pixel will be assigned $BS(P)=1$ as foreground, otherwise $BS(P)=0$ as background, as shown in Fig. 5(a). In the binary switch (BS) map, the foreground clusters correspond to regions which are most likely to contain text information.

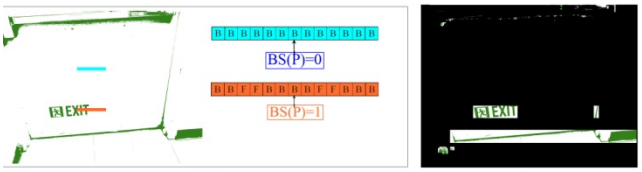

Figure 5: (a) No binary switch occurs within the sliding window in cyan while binary switches are detected within the one in orange; (b) the extracted text regions.

To extract the text regions, an accumulated calculation is developed to locate the rows and columns that belong to the regions in the BS map by Eq. (1), where $H(i)$ and $V(j)$ denote the horizontal projection at row i and vertical projection at column j in the BS map respectively.

$$H(i) = \sum_{j} BS[P(i,j)], \ \text{Row}_i \in \text{TextRegion} \ if \ H(i) > T_H$$
$$V(j) = \sum_{H(i)>T_H} BS[P(i,j)], \ \text{Col}_j \in \text{TextRegion} \ if \ V(j) > T_V \quad (1)$$

The extracted text region is presented in Fig. 5(b). The non-text regions will further be filtered out based on structure of text characters. In the extracted ROIs, text characters with larger font size than their neighbor are often truncated because the extended parts without enough binary switches compared with other text characters. Thus we either extend the text regions to cover the complete boundaries or remove the truncated boundaries.

3.2 Text Localization

To localize text characters and filter out non-text regions from the coarse extracted ROIs, we use a text structure model to identify letters and numbers. This model is based on the fact that each of these characters is shaped by closed edge boundary which has proper size and aspect ratio with no more than two holes (e.g. "A" contains one hole and "8" contains two). Based on the observation that text information usually appears as a string rather than a single character, layout analysis of text strings are taken into account, including the area variance and the neighboring character distance variance of a text string, as shown in Fig. 6. The only assumption is that a text string has at least three characters in uniform color and linear alignment.

To recognize the extracted text information, we employ off-the-shelf OCR software to translate the text information from image into readable text codes. OCR software cannot read the text information directly from indoor sign images without performing text region extraction and text localization. In our experiment, Tesseract and OmniPage OCR software are used for text recognition. Tesseract is open-source without graphical user interface (GUI). OmniPage is commercial software with GUI and performs better than Tesseract for our testing.

Figure 6: In the top row, the inconsistency of $|EF|$ and $|MN|$ leads to separation of two text strings and the arrow is filtered out; in the bottom row, area of M demonstrates inconsistency because it is much larger than its neighbors.

4. EXPERIMENT AND DISCUSSION

We have constructed a database containing 221 images collected from a wide variety of environments to test the performance of the proposed door detection algorithm and text recognition. The database includes both door and non-door images. Door images include doors and elevators with different colors and texture, and doors captured under different viewpoints, illumination changes, and occlusions. Non-door images include door-like objects, such as bookshelves and cabinets. Furthermore, we categorize the database into three groups: *Simple* (57 images), *Medium* (113 images), and *Complex* (51 images), based on the complexity of backgrounds, intensity of deformation, and occlusion, as well as the changes of illuminations and scales.

We evaluated the proposed algorithm with and without performing the function of differentiating doors from door-like convex objects. For images with resolution of 320×240, the proposed algorithm achieve 92.3% true positive rate with a false positive rate of 5.0% without convex object detection. With convex objects detection, the algorithm achieves 89.5% true positive rate with a false positive rate of 2.3%. Table 2 displays the details of detection results for each category with convexity detection and without convexity detection. Some examples of door detection are illustrated in Fig. 7. Our text region extraction

algorithm performs well on extracting text regions from indoor signage and localizing text characters in those regions. OCR is employed on the extracted text regions for text recognition. The last row of Fig. 7 shows that recognized text from OCR as readable codes on the extracted text regions. Of course, if we had input the unprocessed images of the indoor environment directly into the OCR engine, only messy codes would have been produced.

Table 2. Door detection results for groups of "*Simple*", "*Medium*", and "*Complex*"

Data Category	True Positive Rate	False Positive Rate	Convexity Detection
Simple	98.2%	0%	Yes
	98.2%	1.8%	No
Medium	90.5%	3.5%	Yes
	91.4%	6.2%	No
Complex	80.0%	2.0%	Yes
	87.8%	5.9%	No
Average	**89.5%**	**2.3%**	**Yes**
	92.3%	**5.0%**	**No**

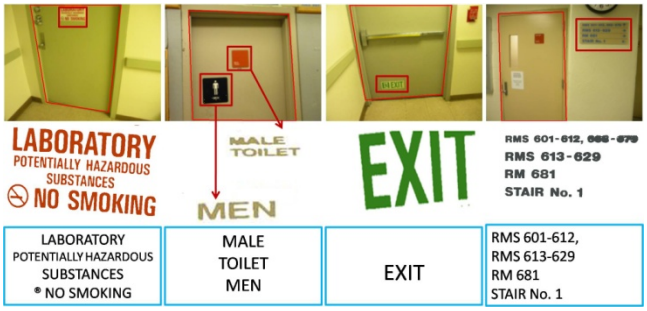

Figure 7: The top row shows the detected doors and regions containing text information; the middle row shows the extracted and binarized text regions; the bottom row shows the text recognition results of OCR in text regions.

As shown in Table 2, from "*Simple*" to "*Complex*", the true positive rate decreases and the false positive rate increases since the complexity of background increases a lot. The convexity detection is effective to lower the false positive rate and maintain the true positive rates of "*Simple*" and "*Medium*". In images with highly complex backgrounds, some background corners adjacent to a door constitute a spurious lateral. So, this door is detected as a convex object and eliminated. However, considering the safety of blind users, the lower false positive rate is more desirable. Text extraction results in a purified background while text localization provides the text positions for OCR. But some background interferences with similar color or size to text characters are still preserved, because the pixel-based algorithms cannot distinguish them from text by high-level structure.

5. Conclusion

We have presented a computer vision-based indoor wayfinding aid to assist blind persons accessing unfamiliar environments by incorporating text information with door detection. A novel and robust door detection algorithm is proposed to detect doors and discriminate other objects with large rectangular shape such as bookshelves and cabinets. In order to help blind people distinguish an office door from a restroom door, we have proposed a new method to extract and recognize text from indoor signage. The recognized text is incorporated with the detected door and further represented to blind persons in audio display.

Our future work will focus on detecting and recognizing more types of indoor objects and icons on signage in addition to text for indoor wayfinding aid to assist blind people travel independently. We will also study the significant human interface issues including auditory output and spatial updating of object location, orientation, and distance. With real-time updates, blind users will be able to better use spatial memory to understand the surrounding environment.

6. ACKNOWLEDGMENTS

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

7. REFERENCES

[1] Anguelov D., Koller D., Parker E., and Thrun S., Detecting and modeling doors with mobile robots. In *Proceedings of the IEEE International Conference on Robotics and Automation*, 2004.

[2] Chen Z. and Birchfield S., Visual Detection of Lintel-Occluded Doors from a Single Image. *IEEE Computer Society Workshop on Visual Localization for Mobile Platforms*, 2008.

[3] Hensler J., Blaich M., and Bittel O., Real-time Door Detection Based on AdaBoost Learning Algorithm. *International Conference on Research and Education in Robotics, Eurobot* 2009.

[4] T. Kasar, J. Kumar and A. G. Ramakrishnan. "Font and Background Color Independent Text Binarization". *Second International Workshop on Camera-Based Document Analysis and Recognition*, 2007.

[5] Liu, C., Wang, C., and Dai, R.,Text Detection in Images Based on Unsupervised Classification of Edge-based Features, *International Conference on Document Analysis and Recognition*, 2005.

[6] Liu, Q., Jung, C., and Moon, Y.,Text Segmentation Based on Stroke Filter, *Proceedings of International Conference on Multimedia*, 2006.

[7] Murillo A. C., Kosecka J., Guerrero J. J., and Sagues C., Visual door detection integrating appearance and shape cues. *Robotics and Autonomous Systems*, 2008.

[8] Shivakumara, P., Huang, W., and Tan, C., An Efficient Edge based Technique for Text Detection in Video Frames, *The Eighth IAPR Workshop on Document Analysis Systems*, 2008.

[9] Stoeter S., Mauff F., and Papanikolopoulos N., Realtime door detection in cluttered environments. In *Proceedings of the 15th IEEE International Symposium on Intelligent Control*, 2000.

[10] Wong, E., and Chen, M., A new robust algorithm for video text extraction, *Pattern Recognition* 36, 2003.

[11] Yang X. and Tian Y., Robust Door Detection in Unfamiliar Environments by Combining Edge and Corner Features. *IEEE Computer Society Workshop on Computer Vision Applications for the Visually Impaired*, 2010.