Machine Vision and Applications (2013) 24:521–535
DOI 10.1007/s00138-012-0431-7

ORIGINAL PAPER

Toward a computer vision-based wayfinding aid for blind persons to access unfamiliar indoor environments

YingLi Tian · Xiaodong Yang · Chucai Yi · Aries Arditi

Received: 23 May 2011 / Revised: 10 March 2012 / Accepted: 23 April 2012 / Published online: 24 May 2012

Abstract Independent travel is a well-known challenge for blind and visually impaired persons. In this paper, we propose a proof-of-concept computer vision-based wayfinding aid for blind people to independently access unfamiliar indoor environments. In order to find different rooms (e.g. an office, a laboratory, or a bathroom) and other building amenities (e.g. an exit or an elevator), we incorporate object detection with text recognition. First, we develop a robust and efficient algorithm to detect doors, elevators, and cabinets based on their general geometric shape, by combining edges and corners. The algorithm is general enough to handle large intra-class variations of objects with different appearances among different indoor environments, as well as small inter-class differences between different objects such as doors and door-like cabinets. Next, to distinguish intra-class objects (e.g. an office door from a bathroom door), we extract and recognize text information associated with the detected objects. For text recognition, we first extract text regions from signs with multiple colors and possibly complex backgrounds, and then apply character localization and topological analysis to filter out background interference. The extracted text is recognized using off-the-shelf optical character recognition software products. The object type, orientation, location, and text information are presented to the blind traveler as speech.

Keywords Indoor wayfinding · Computer vision · Object detection · Text extraction · Optical character recognition (OCR) · Blind/visually impaired persons

Y. Tian (✉)
Electrical Engineering Department, The City College,
and Graduate Center, City University of New York,
New York, NY 10031, USA
e-mail: ytian@ccny.cuny.edu

X. Yang
Electrical Engineering Department, The City College,
City University of New York, New York, NY 10031, USA
e-mail: xyang02@ccny.cuny.edu

C. Yi
The Graduate Center, City University of New York,
New York, NY 10036, USA
e-mail: cyi@gc.cuny.edu

A. Arditi
Visibility Metrics LLC, Chappaqua, NY 10514, USA
e-mail: arditi@visibilitymetrics.com

1 Introduction

Robust and efficient indoor object detection can help people with severe vision impairment to independently access unfamiliar indoor environments and avoid dangers [3]. While GPS-guided electronic wayfinding aids show much promise in outdoor environments, there are few indoor orientation and navigation aids [15,28]. Computer vision technology in principle has the potential to assist blind individuals to independently access, understand, and explore such environments. Yet it remains a challenge for the following four reasons. First, there are large intra-class variations of appearance and design of objects in different architectural environments. Second, there are relatively small inter-class variations of different object models. Third, relative to richly textured and colored objects in natural scene or outdoor environments, most indoor objects are man made and have little texture. Feature descriptors which work well for outdoor environments may not effectively describe indoor objects. Finally, object with large view variations and often only parts of object (within the field of view) are captured when a blind user moves. An effective indoor wayfinding aid should handle object occlusion and view variations.

In our approach, we exploit the fact that context information (including signage) plays an important role in navigation and wayfinding for sighted persons. Signage is particularly important for discriminating between similar objects in indoor environments such as elevators, bathrooms, exits, and office doors. 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.

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 proof-of-concept navigation and wayfinding prototype system by using a single camera to detect and recognize doors and elevators. The system can further detect text signage associated with the detected object. Our object detection method is based on general geometric shape by analyzing configuration of edges and corners. The proposed algorithm is robust, efficient, and generic enough to handle large intra-class variations of object appearances across different environments, as well as small inter-class variations of different objects such as doors and door-like shape cabinets. In some cases, text information from the signage associated with the detected objects is extracted in order to distinguish more subtle differences, such as distinguishing an office door from a bathroom door. Note that due to the Americans with Disabilities Act and other legislation, most doors in public accommodations are now required to be labeled with appropriate signage that includes Braille. Braille is difficult for blind users to locate haptically, however, and is of limited value in wayfinding. See [1] for a discussion of these issues. Our text extraction method is robust to indoor signage with multiple colors and complex backgrounds. The extracted text is then recognized by using off-the-shelf optical character recognition (OCR) software. The object type, relative position, and text information are presented as speech for blind travelers.

2 Related work

2.1 Computer vision-based technologies for blind persons

There have been many previous efforts to apply computer vision-based technologies to aid blind persons [5,7,9,14,18,25,32,33,40,44]. A survey about recent research on wearable obstacle avoidance electronic travel aids for people with visual impairments can be found in paper [10]. The vOICe vision technology offers the experience of live camera views through proprietary image-to-sound renderings [40]. The Smith-Kettlewell Eye Research Institute developed a series of camera phone-based technology tools and methods for understanding, assessment, and rehabilitation of blindness and visual impairment, including text detection and location [33], crosswalk identification and location [18], and a wayfinding system based on machine-readable codes applied to the environment [25]. Chen and Yuille [7] developed an AdaBoost learning-based algorithm to detect and read text in natural scenes. Zandifar et al. [44] used one head-mounted camera together with existing OCR techniques to detect and recognize text in outdoor environments and then convert the text to speech. Everingham et al. conceived of a wearable mobility aid for people with low vision for outdoor scene classification in a Markov random field model framework based on color information [14]. Although many efforts of finding ways to apply vision technology helping blind people understand their surroundings have been made, no approach as yet has been widely or successfully adopted. Recently, our efforts for developing robust algorithms of indoor object detection in unfamiliar environments demonstrated promising results for indoor wayfinding and navigation applications [9,37,43].

2.2 Indoor object detection

Object detection and recognition is a fundamental component of wayfinding. The human visual system is very powerful, selective, robust, and fast [21] and can discriminate on the order of tens of thousands of different object categories [4] In stark contrast to the exquisite capabilities of human vision, it is extremely difficult to build robust and selective computer vision algorithms for general object detection and recognition. In this paper, we simplify the problem by focusing on object detection for indoor wayfinding and navigation for the following reasons: (1) lighting and luminance are relatively stable in indoor environments; (2) recognition and localization of a small and well-defined set of objects such as doors, elevators, bookshelf/cabinet, and signage provide a smaller subset of information likely to be particularly useful for blind travelers; (3) indoor environments are structured by artificial objects with relatively standard size, shape, and location; and (4) indoor environments are relatively safe for blind users who participate in our research.

Among indoor objects, doors are important landmarks for wayfinding and navigation and play a significant role in providing transition points between separated spaces as well as entrance and exit information. Therefore, reliable and efficient door detection is a key component of an effective indoor wayfinding aid. Most existing door detection approaches for robot navigation employ laser range finders, sonar, or stereo vision to obtain distance data to refine the detection results from cameras [2,17,20,35]. However, the problems of portability, high-power, and high cost in these systems with complex and multiple sensors, limit their feasibility in wayfinding systems for visually impaired people. To reduce the cost and complexity of the device and enhance the portability, we use a single camera in our system. A few algorithms using

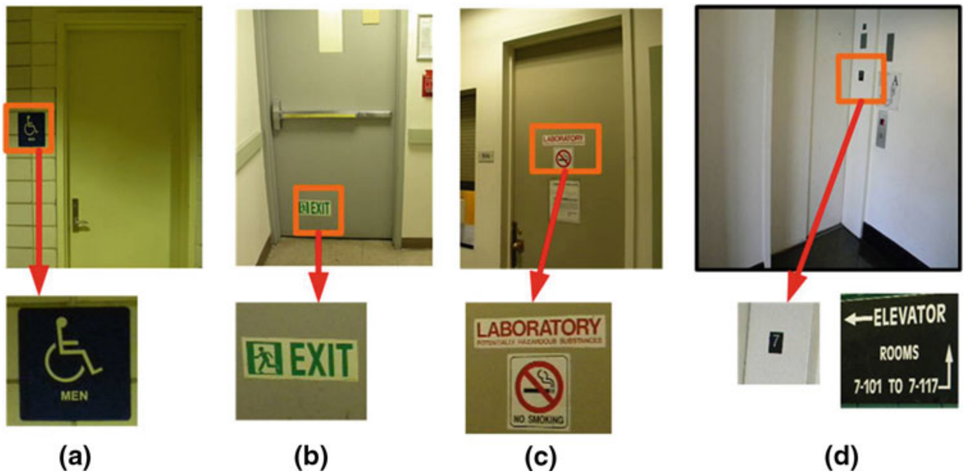


Fig. 1 Typical indoor objects (*top row*) and their associated contextual information (*bottom row*). **a** A bathroom, **b** an exit, **c** a laboratory, and **d** an elevator

monocular visual information have been developed for door detection [8,26,27]. Chen and Birchfield [8] trained an AdaBoost classifier to detect doors by combining the features of pairs of vertical lines, concavity, gap between the door and floor, color, texture, kick plate, and vanishing point. However, in practice, some of these features (e.g. a perceptible gap below a door and the floor and a kick plate) are not always present in different instances of a door. In [27], an algorithm is described that detects the doors of one building, where all the doors are of similar color, by using color and shape features. It would fail, however, if the colors of the doors varied. Munoz-Salinas et al. [26] developed a doorframe model-based algorithm by using Hough Transform to extract frame segments. They subsequently use fuzzy logic to analyze relationships between the segments. Their algorithm, however, cannot discriminate doors from other large rectangular objects, such as bookshelves, cabinets, and cupboards.

To overcome the above limitations, we have developed an image-based door detection algorithm to detect doors which may have variable appearance, by adopting a very general geometric door model that utilizes the general and stable features of doors—edges and corners. Furthermore, our proposed algorithm is able to differentiate doors from other objects with door-like shape and size by analyzing geometric cues that are apparent when the door is inset within a wall, such as is especially common with elevator doors (see Fig. 4a, b for illustration). Our detection results (see below) demonstrate that our door detection algorithm is general and robust to different environments with a wide variety of color, texture, occlusions, illumination, scales, and viewpoints.

2.3 Context information extraction and recognition

Context information obviously plays an important role in human object identification and detection [24,30,31,38]. Paletta and Greindl [31], for example, extracted context from simple features to determine regions of interest in a multi-stage object detection process. Luo et al. [24] developed a spatial context-aware object detection system to improve the accuracy of natural object detection. Divvala et al. [12] investigated the different types of contextual information on a standard, highly regarded test set (the PASCAL VOC 2008) by incorporating contextual information into a post-process which re-scores detection hypotheses based on their coincidence with the various context cues.

Context information includes many kinds of visual features. In our method, we focus on finding and extracting a particularly useful kind of feature that is particularly amenable to computer vision analysis – text from signage. In recent years, there have been many efforts to identify and extract text regions from images. Dubey [13] used statistics of the frequency of vertical stroke occurrences, while Dinh et al. [11] focused on the consistency of stroke width. Neither of these algorithms performs well when the number of continuous text characters is small. Dense intensity variation based on the edge map is another potentially useful texture-like feature of text characters. Shivakumara et al. [34] worked on different characteristics of edge detectors. Wan et al. [41] performed edge classification by corner information. Liu et al. [22] described edge features by defining six characteristic values involved in four dominant directions of edge detection. Different sliding windows are defined to transverse the whole image to extract text information. In addition to dense intensity variations, Liu et al. [23] brought in texture features of text characters. Wong et al. [42] applied maximum gradient difference (MGD) to detect potential line segments that constitute the text regions. The ridge points with local extreme intensity variation were employed in [39] to describe text strokes at a high resolution and text orientations at a low resolution. However, the calculation of both MGD and ridge points was sensitive to background noise.

We propose a new and robust algorithm to extract text from signage. Both structural and topological features of text characters in the same text string are employed to refine the

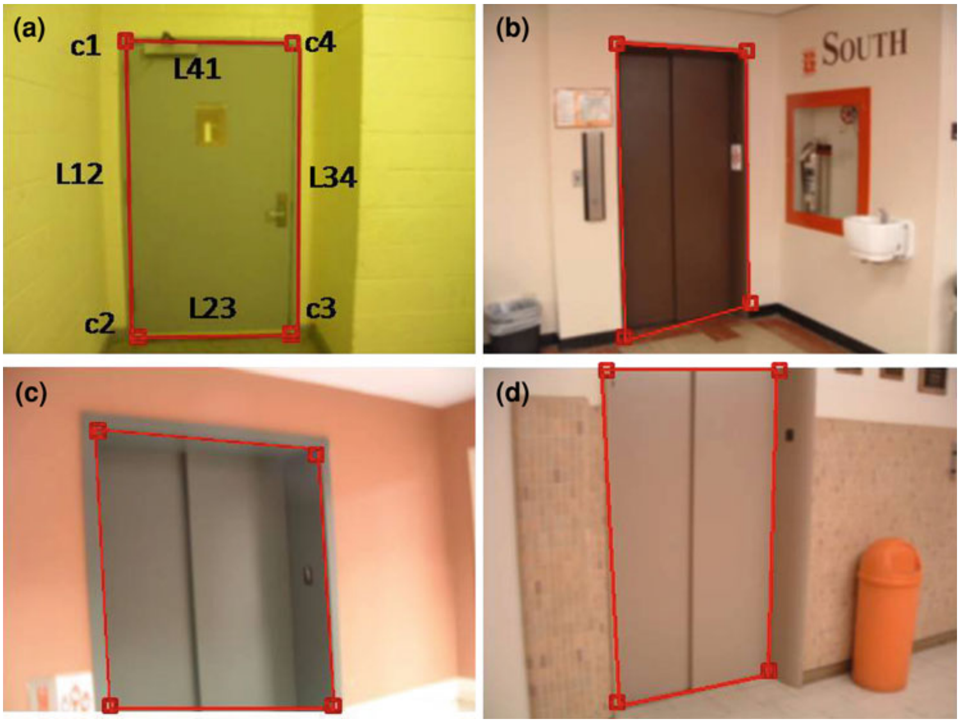


Fig. 2 Geometric door model: **a** ideal 'canonical' view, **b** with perspective effects. **c**–**d** with occlusion and perspective effects

detection and localization of text region. More details can be found in the paper [36].

3 General indoor object detection algorithm

3.1 Door detection

To robustly detect doors in different environments, we use edges and corners to model the geometric shape of a door frame, which contains two horizontal lines and two vertical lines between four corners. We extract edges using the Canny edge detector [6] and corners through the corner detector described in [16]. We further propose a matching method that combines edges and corners rather than directly detecting door frames. This matching process can avoid some pitfalls of existing line detection algorithms, such as missing start and end points, sensitivity to parameters, and unwanted merging or splitting of lines. In addition, we use perspective geometry to estimate the relative position of the detected door for the blind user. Furthermore, in combination with information about neighboring lateral areas, the door detection algorithm is able to indirectly obtain the depth information of the doorframe, which can be used to differentiate doors from other indoor objects with door-like shape and size.

3.1.1 Geometric door model

Our geometric model of a door consists of four corners and four lines. Figure 2a depicts the ideal image view without occlusion or perspective. Figure 2b shows the deformed geometric shape caused by perspective. In both cases, the four door corners are all visible. Often, however, due to perspective and occlusion, only a part of a door is captured by a wearable camera, especially for visually impaired users who cannot aim the camera to "frame" the door. As shown in Fig. 2c, d, each occluded vertical line can form a corner with the horizontal axis of an image. In our algorithm, we assume the following about the image: (1) At least two door corners are visible. (2) Both vertical lines of a door frame are visible. (3) Vertical lines of a door frame are nearly perpendicular to the horizontal axis of an image. (4) A door has at least a certain width and length (to be defined in the next section). These assumptions are easy to achieve in practice. The geometric door model is robust to variations in color, texture, occlusion and door status without any other requirements. For instance, the color of a door can be similar or different from that of an adjacent wall (provided the door frame is contrasting in color). The surface of a door can be highly textured or not. The status of a door may be closed, open, or partially open.

3.1.2 Door-corner candidates

Edges and corners are insensitive to variations in color, viewpoints, scales, and illumination. Our model employs a combination of edges and corners to represent a door, which generalizes to doors in different environments. We first apply pre-processing (i.e. down-sampling, Gaussian smoothing) to eliminate noise for reducing corners detected from highly textured surrounding walls. Then the Canny edge detection is performed on the down-sampled and smoothed image to generate a binary edge map. Contours are extracted from the edge map by connecting the gaps between close endpoints of edges. Next, corners are extracted from contours using the method of He and Yung [16]. In the geometric door model, each line terminated by the image border corresponds to an open contour. Therefore, four corners of the geometric door model (see Fig. 2) can always be extracted even when only the upper (or lower) part of the door is captured.

Door-corner candidates can be grouped from the detected corners in the image based on the relationship of four corners of the doorframe in the geometric door model. As shown in Fig. 2a, for any combination of four corners, the top left corner is denoted C_1, other three corners are named as C_2, C_3, and C_4 in counterclockwise direction. The coordinate of corner C_i is (x_i, y_i). L_{12} is the line connecting C_1 and C_2, similarly for L_{23}, L_{34}, and L_{41}. The ratio Siz_{ij} between the length of L_{ij} and the length of the diagonal *DI* of an image, and the direction of L_{ij} corresponding to the horizontal axis of an image Dir_{ij} can be used to obtain the door-corner candidacy from detected corners.

The combination of four corners will be selected as a door-corner candidate if the following requirements are satisfied:

(1) A door in an image has a certain height and width. So, Siz_{12} and Siz_{34} should be within a range:

$$\begin{aligned} HeightThreshL &< Siz_{12}, Siz_{34} \\ &< HeightThreshH \\ WidthThreshL &< Siz_{23}, Siz_{41} \\ &< WidthThreshH \end{aligned}$$

(2) Ideally, L_{41} and L_{23} should be parallel with the horizontal axis of an image. Due to perspective, L_{41} and L_{23} will form an angle with the horizontal axis. Therefore, Dir_{41} and Dir_{23} cannot be too large:

$$Dir_{23}, Dir_{41} < DirectionThreshL$$

(3) Vertical lines of a door frame are almost perpendicular to the horizontal axis of an image. So, Dir_{12} and Dir_{34} should be large enough:

$$Dir_{12}, Dir_{34} > DirectionThreshH$$

(4) Vertical lines of a door frame should be approximately parallel with each other:

$$|Dir_{12} - Dir_{34}| < ParallelThresh$$

(5) The ratio between height and width of a doorframe should be within a range:

$$\begin{aligned} HWThreshL &< (Siz_{12} + Siz_{34}) / (Siz_{23} + Siz_{41}) \\ &< HWThreshH \end{aligned}$$

Based on the camera configuration and door size, we set *HeightThreshL = 0.5* and *HeightThreshH = 0.9*; *WidthThresh L = 0.1* and *WidthThreshH* = 0.8; *DirectionThreshL = 35* and *DirectionThreshH = 80*; *ParallelThresh* = 6; *RatioThreshL = 2.0* and *RatioThreshH* = 3.0 in our system. The four-corner groups that satisfy the above rules are marked as door-corner candidates, which are subsequently tested by the following matching procedure.

3.1.3 Door detection by matching edges and door-corner candidates

Each door-corner candidate represents an abstract geometric frame. To determine whether a door-corner candidate represents a real doorframe, we check if there are matching edges of the door frame between corners of each door-corner candidate based on our geometric door model. As shown in Fig. 2a, C_i and C_j are two corners of a door-corner candidate. We first create a matching mask (i.e. the gray area) by expanding the line connecting C_i and C_j with a $N \times N$ window. The "fill-ratio" FR_{ij} of corner C_i and C_j is defined as the ratio of the overlap between the detected edge pixels falling in the matching mask and the line connecting the two corners. Combining the edge map and door-corner candidates, we can obtain a "fill-ratio" vector $[FR_{12}, FR_{23}, FR_{34}, FR_{41}]$ for each door-corner candidate. If each element of a "fill-ratio" vector is larger than *FRThreshL* and the average value of four elements is larger than *FRThreshH*, then this door-corner candidate corresponds to a real door in the image. If there is more than one door-corner candidate with matching edges of the same doorframe, they will be merged as one detected door.

In practice, if *FRThreshL* or *FRThreshH* is too large, the false-positive rate will be low, but the true-positive rate will decrease; if *FRThreshL* or *FRThreshH* is too small, the true-positive rate will be high, but the false-positive rate will increase. In order to reduce the false-positive rate and increase the detection rate simultaneously, we initialize *FRThreshL* and *FRThreshH* by two relatively low values. Then, the detection result is checked: if only one door-corner candidate is matched, it is the detection result; if more than one door-corner candidate is matched, then *FRThreshL* and *FRThreshH* with relatively high values are used to re-match the matched door-corner candidates. In our experiments, the increased thresholds are effective in eliminating spurious detection results. In our experiments, parameter values were

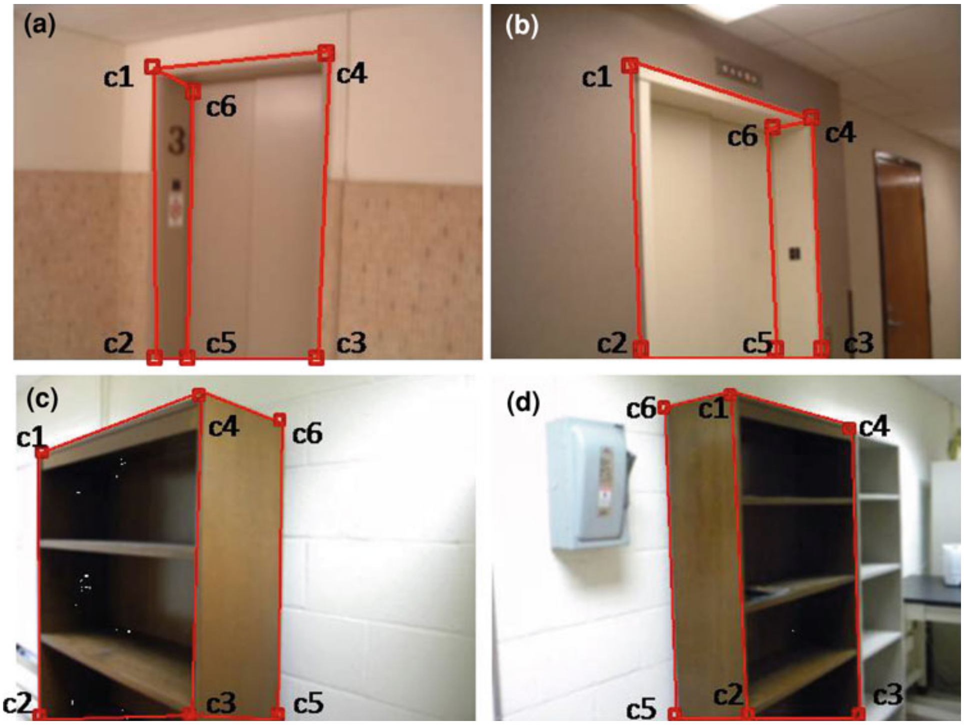


Fig. 3 Inset and protruding models. **a** Inset object: door frame located on user's right side and lateral face located on the doorframe's left side. **b** Protruding object: frame located on the user's left side and lateral face located on the frame's right side. **c** Protruding object: frame located on the user's right side and lateral face located on the frame's right side. **d** Protruding object: frame located on the left side; lateral face located on the left side

$FRThreshL = 0.6$ and $FRThreshH = 0.85$ in the first matching process. They are set to 0.87 and 0.90, respectively, in the re-matching process.

3.2 Determining relative position of a detected door

For an indoor wayfinding device to assist blind or visually impaired persons, we not only need to recognize the presence of a door but also the door's position relative to the user. Figure 11a depicts the scenario that a door is located on the left side with respect to the camera. Figure 11b demonstrates a door located in front of the observer and Fig. 11c demonstrates a door located on the right side. The angle between L_{41} and the horizontal axis (see Fig. 2a) can be used to determine the relative position of a door.

3.3 Inset and protrude object detection

Depth information plays an important role in indoor object recognition, as well as in robotics, scene understanding, and 3-D reconstruction. In most cases, doors are recessed or inset into a wall, especially the doors of elevators. Other large objects with door-like shape and size, such as bookshelves and cabinets, protrude in relief from a wall. In order to distinguish inset (e.g. doors and elevators) from protruding (e.g. bookshelves, cabinets) objects, we propose a novel and straightforward algorithm by utilizing information from faces formed by an inset or protrusion to obtain the depth information with respect to the wall.

Table 1 Protrusion and inset determination combining the position of a frame and the position of a lateral face

Position of a frame	Position of a lateral face	Inset or protrusion
Right	Left	Inset (Fig. 3a)
Left	Right	Inset (Fig. 3b)
Right	Right	Protrusion (Fig. 3c)
Left	Left	Protrusion (Fig. 3c)

Due to the different geometric characteristics and different relative positions, the different positions of lateral faces with respect to the detected object indicate door-like inset objects (assumed to be doors) or protruding objects (assumed to be other furnishings such as bookshelves, as shown in Fig. 3. In Fig. 3a, since $L_{12}<L_{34}$, the elevator door is determined to be located on the right side with respect to the user. Furthermore, the elevator door, an inset object, presents 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. 3c, since $L_{12}<L_{34}$, the bookshelf is identified as located on the right side and as a protruding object, its lateral $(C_4 - C_3 - C_5 - C_6)$ being on the right side of the frame $(C_1 - C_2 - C_3 - C_4)$. Similar relations can be found in Fig. 3b, d. Therefore, combining the position of a frame and the position of a lateral, we can determine the inset or relief of a frame-like object, 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.

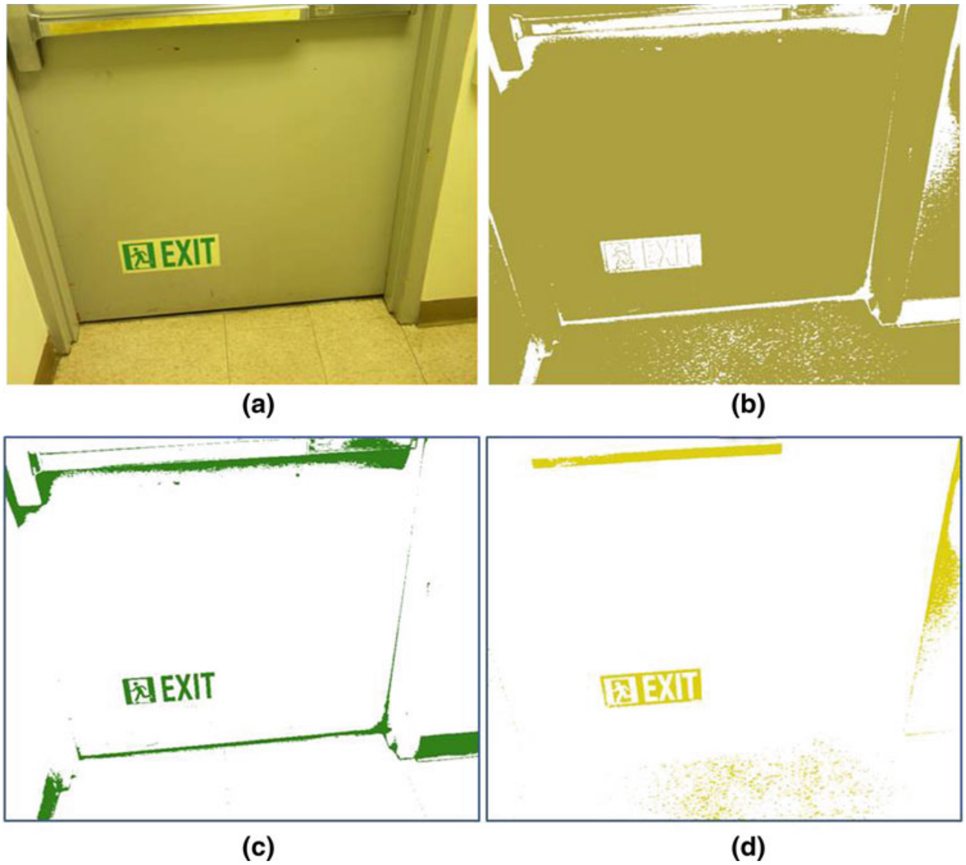


Fig. 4 Three color layers are extracted from an exit door. **a** Original image. **b**–**d** Different color layers extracted by color decomposition

4 Text extraction, localization, and recognition algorithm

4.1 Text region extraction

4.1.1 Color decomposition

To extract text information from indoor signage with complex backgrounds, we have designed a robust algorithm to extract the coarse regions of interest (ROI) that are most 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 clean background. To obtain fewer layers, we apply color quantization to reduce the number of colors in the original image, but keep the essential information for text extraction and segmentation. Inspired by Nikolaou et al. [29], first, we perform edge detection to create the edge map. Second, to avoid the drastic color variations around edge pixels, only non-edge pixels are sampled for the next step. Third, to group pixels with similar colors together, seeds are randomly selected from the sampled pixels to cluster the pixels with similar colors to the corresponding seed pixels. Fourth, considering the mean color value of each color cluster as an updated seed, a mean shift-based clustering algorithm [19], a nonparametric clustering technique which does not require prior knowledge of the number of clusters, is applied iteratively to group clusters with similar color together to produce larger clusters, each of which corresponds to a color layer. Each color layer is a binary image with one foreground color and white background color. Figure 4 shows an example of an exit door image with three color layers after color decomposition.

4.1.2 Text region extraction based on density analysis of binary transitions

After color decomposition, we obtain the color layer images that contain regions with similar colors. Now, we extract text regions based on the structure of text characters. Text characters that are composed of strokes and major arcs are well aligned along the direction of whatever text string they belong to. For each text string, well-structured characters result in a high frequency of regular intensity variations. Since each color layer is a binary image, the regular intensity variations between the text color and the background color are represented as binary transitions in the highly textured regions. Generally, text strings on indoor signage have layouts with a dominant horizontal direction with respect to the image

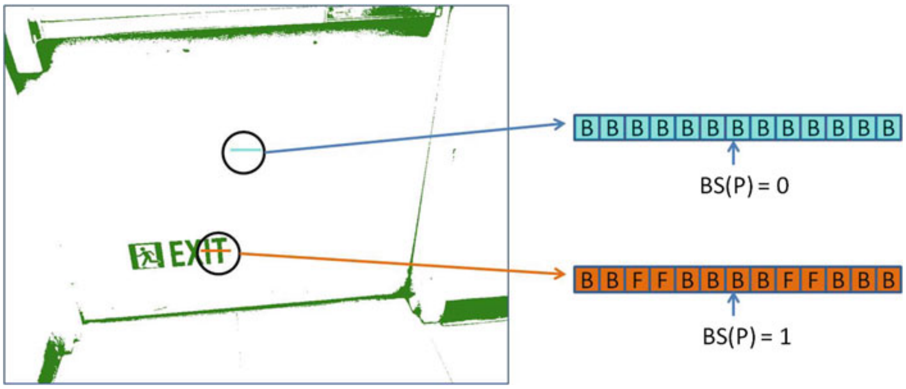


Fig. 5 No binary transition occurs within the sliding window in *blue*, while binary transitions are detected within the one in *orange* (color figure online)

frame. We apply a 1 × N sliding window to calculate the binary transitions around the center pixel of the window P. If there are no less than two binary transitions occurring inside the sliding window, the corresponding central pixel will be assigned *BS(P)=1*, otherwise *BS(P)=0*, as shown in Fig. 5. Thus, a map of binary transitions for each color layer is created to accurately represent the regional intensity variation. In the binary transitions map, the clusters constituted by value 1 correspond to higher frequencies of binary transitions, which mean that they are most likely regions containing text information.

To extract text regions with high frequencies of binary transitions, an accumulation calculation is used to locate the rows and columns that belong to the regions in the BS map by Eqs. (1) and (2). Let $H(i)$ and $V(j)$ indicate the horizontal projection at row i and vertical projection at column j in the BS map, respectively. If $H(i)$ is greater than a threshold, row i is part of a text region. Accumulation is then performed vertically in the regions constituted by extracted rows; if $V(j)$ is greater than a threshold, the column j located in extracted rows are also part of the text region.

$$H(i) = \sum_{j} BS[P(i,j)], \mathrm{Row}_i \in \mathrm{TextRegion if} H(i) > T_H \tag{1}$$

$$V(j) = \sum_{H(i)>T_H} BS[P(i,j)], \mathrm{Col}_j \in \mathrm{TextRegion if} V(j) > T_V \tag{2}$$

Therefore, the regions that are very likely to contain text information are extracted as ROIs. Note that in this stage, the extracted ROIs still contain some highly textured non-text regions. These non-text regions will further be filtered out based on structure of text characters (see details in Sect. 4.2).

4.1.3 Text region extension

In the extracted ROIs, text characters that are larger than their neighbors (as shown in Fig. 6) might be truncated, because the top parts have fewer binary transitions than lower text characters. In order to acquire regions with fully formed text characters, we check boundaries of each component in an extended text region. As shown in Fig. 6, if the complete boundaries could be covered by a restricted extension of the region's size, the stretched region will be deemed as a new ROI.

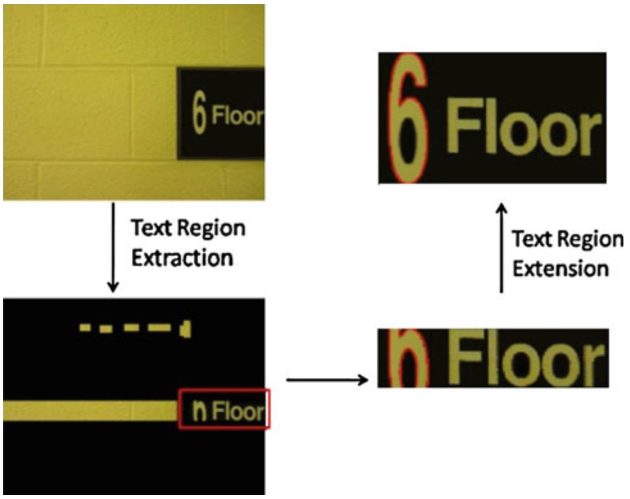


Fig. 6 Region extension by recovering the truncated boundary denoted in *red* (color figure online)

4.2 Text localization

4.2.1 Primary localization of text characters

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 contains no more than two holes (e.g. "A" contains one hole and "8" contains two holes).

As in the algorithm proposed by Kasar et al. [19], the edge map of the text regions is first computed. The edge bounding box is then obtained by performing connected component labeling. Several rules are defined to filter out obvious non-

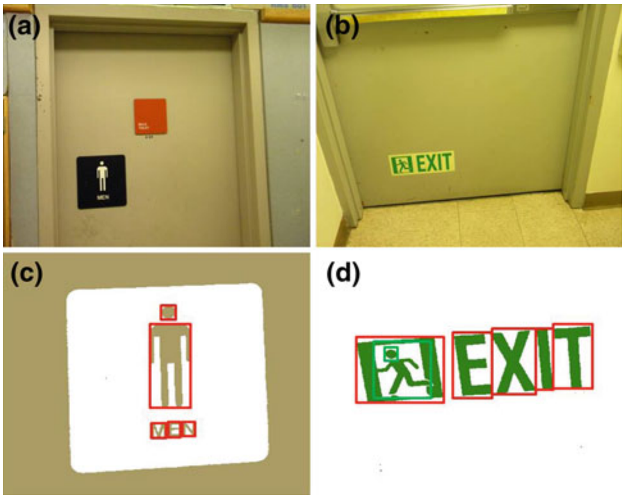


Fig. 7 **a, b** Original images. **c, d** Bounding boxes at the extracted regions. The *green boxes* are the non-text regions which will be filtered out, leaving only *red boxes* for topological analysis (color figure online)

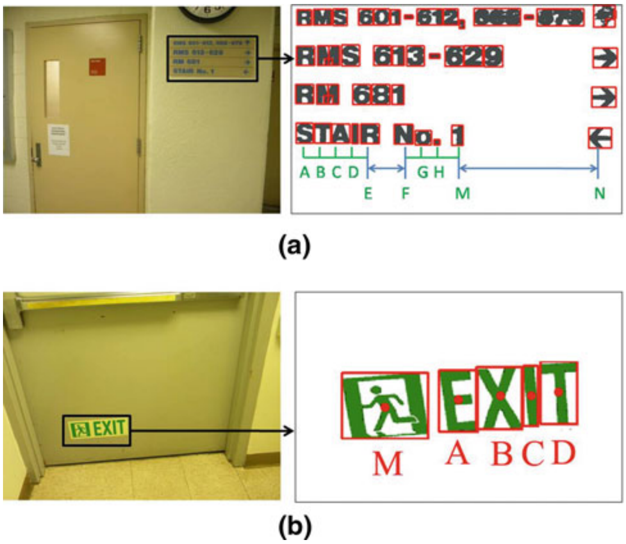


Fig. 8 **a** $|EF|$ and $|MN|$ demonstrate inconsistency compared to the neighboring distances, so two text strings are separated and the *arrow* is filtered out. **b** Area of M demonstrates inconsistency, because it is much larger than its neighbors

text bounding boxes, including the size, aspect ratio, and the number of nested bounding boxes. Figure 7 shows two examples of text character localization. The green box in Fig. 7d is a non-text region which will be filtered out by aspect ratio of the bounding box, and red boxes represent the bounding box of text characters for further topological analysis.

4.2.2 Text localization refinement by topological analysis

To further filter out non-text bounding boxes, topological features from text strings are taken into account. We assume that a text string contains at least three three characters, as shown in Fig. 8. Based on the observation that the distances and areas between neighboring characters are highly consistent in text strings, we first measure the distance variation among three neighboring bounding boxes in accordance with their central points, as shown in Fig. 8a. In addition to filtering out the non-text bounding boxes with inconsistent inter-letter or inter-word distances, we can also categorize independent text strings. Moreover, the area variation is calculated among the three neighboring bounding boxes by the foreground area ratio of each pair of bounding boxes, as shown in Fig. 8b. If one of them is much larger or smaller than the two neighbors that have similar foreground areas, it will be filtered out. The topological analysis results in purified text strings on cleaner backgrounds.

4.3 Text recognition

To recognize the extracted text information, we employ off-the-shelf OCR software to translate the text images into character codes. Currently available commercial OCR is designed for scanned documents with large regions of text characters in perfect alignment in a clean background. For indoor sign images, OCR software cannot read the text information without performing text region extraction and text localization. In our experiments, Tesseract and OmniPage OCR software are used for text recognition. Tesseract is open-source without graphical user interface (GUI). OmniPage is a commercial software with GUI and performed better than Tesseract in our testing.

5 System and interface design

5.1 Indoor wayfinding system design

The computer vision-based indoor wayfinding aid for blind persons integrates a camera, a microphone, a portable computer, and a speaker connected by Bluetooth for audio description of objects identified. A mini-camera mounted on sunglasses is used to capture video of the environment for computer vision-based object class recognition. The presence of environmental objects (class, location, etc.) is described to the blind user by verbal display with minimal distraction to the user's hearing sense. The user can control the system by speech input via microphone.

Interface design: In order to interact with the blind user, the aid provides function selection and system control through speech commands input from a microphone. As shown in Fig. 9, the basic interface design includes "Basic Functions" and 'High Priority Commands", which will be refined based on users' evaluation and feedback.

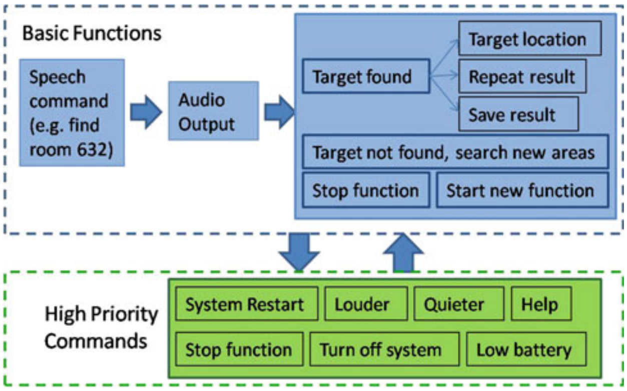


Fig. 9 Basic interface design for the proposed computer vision-based indoor wayfinding system by using speech commands. The high priority commands can be used at any time to overwrite the basic functions

Basic functions: The blind user will speak out the destination or target object for which he/she is looking as the input of the indoor wayfinding aid (e.g. "find exit door"). The object detection results will be presented to the blind user as speech outputs including "Target found", "Target not found", "Stop function", "Start a new function", etc. For the "Target found" function, the next level of functions includes "Target location", to announce the location and orientation of the target, "Repeat", to repeat the detected result, and "Save result", to save the object image with the associated context information in the computer.

High priority system configuration: The system configuration can be set by a number of high priority speech commands such as "System Restart", "Turn off system", "Abort Current Task", speaker volume and speed control commands (e.g. "Louder", "Quieter", "Slower", "Faster"), and "Help". The commands with high priority can be used at any time. The user may verbally request "Help", and the indoor wayfinding system will respond with which options are available within the current function. To protect privacy and minimize masking environmental sounds, bone conduction earphones or small wireless Bluetooth speakers can be used. The system will check battery level and send out an audio warning when the battery level is low.

5.2 Audio output

For the audio display, we use operating system speech facilities, which are standard in modern portable computer systems (and smartphones). Currently, we use the Microsoft Speech software development kit (SDK), which conveniently supports imported script files. Many configuration options are available including speech rate, and volume and voice gender according to user preference. More studies are needed (and planned) on how best to describe complex indoor environments in compact yet effective verbal terms.

5.3 System prototype implementation

As shown in Fig. 15, we have developed a prototype system that implements portions of the design described in this paper. Our research on the interface study is ongoing. The hardware of the prototype system includes a Logitech web camera with auto focus which is attached on a pair of sunglasses to capture images. The camera is connected to an HP mini laptop by a USB connection. The HP mini laptop processes the computation. In order to avoid serious blocking or distracting the hearing of blind people, a sense that they rely upon heavily, we choose a wireless BlueTooth earpiece for presenting detection results as speech outputs to the blind travelers. For speech input and output, we employ Microsoft Speech SDK in our system. For text recognition, we evaluate two OCR engines, *Tesseract* and *Nuance OmniPage*. *OmniPage* demonstrates better performance in most cases, but it is a commercial software without open source codes and expensive. *Tesseract* is an open-source OCR engine that can be more conveniently integrated into systems. We integrate *Tesseract* in our implementation.

6 Experimental results and discussions

6.1 Database for indoor object detection

We constructed a database containing 221 images collected from a wide variety of environments by static cameras to test the performance of the proposed door detection and text recognition algorithms. The database includes both door and non-door images. Door images include doors and elevators with different colors and texture, and doors captured from different viewpoints, illumination conditions, and occlusions, as well as open and glass doors. Non-door images include door-like objects, such as bookshelves and cabinets.

To evaluate the proposed methods, we first categorized 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 changes of illumination and scale. To test the accuracy of door position detection, we further regroup the door images in the database based on the views as: *Left, Frontal, and Right*. There are a total of 209 door images with 36 of *Left*, 141 of *Frontal*, and 32 of *Right*.

6.2 Experimental results

6.2.1 Door detection

We evaluated the proposed algorithm with and without performing the function of differentiating doors from door-like protruding objects. For images with resolution of 320 × 240,

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

Data category	True positive rate (%)	False positive rate (%)	Protrusion 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
Total	89.5	2.3	Yes
	92.3	5.0	No

the proposed algorithm achieved 92.3 % true-positive rate with a false-positive rate of 5.0 % without protruding object detection. With protruding object 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 protrusion detection and without protrusion detection. Some examples of door detection are illustrated in Fig. 10.

6.2.2 *Door position detection*

To determine the relative position of a detected door, we classify the relative position as *Left*, *Front*, and *Right* (see examples in Fig. 11). The relative position detection rate for *Left*, *Front*, and *Right* are 97.2, 95.0, and 93.8 %, respectively. So, the method proposed in this paper appears to be highly reliable for inferring the relative position of a detected door with respect to a user.

6.2.3 *Door signage recognition*

Text region extraction is very important for recognizing indoor signage and is of course necessary before OCR can be performed for text recognition. Some example results are demonstrated in Fig. 12. The first row of Fig. 12 shows the detected door and signage regions. The second row displays the binarized signage. The last row of Fig. 12 displays that recognized text from OCR as readable codes on the extracted and binarized text regions. Of course, if we had input the unprocessed images of the indoor environment directly into the OCR engine, only messy codes, if any, would have been produced.

We quantitatively evaluated the proposed text localization method against the Robust Reading Dataset [45] from International Conference on Document Analysis and Recognition (ICDAR) 2003. In our testing, we selected 420 images which are compatible with the assumption that a text string contains at least three characters with relatively uniform color. To evaluate the performance, we calculate the precision, which is the ratio of area of the successfully extracted text regions to area of the whole detected regions. Our method achieves the state of the art with an average precision of 71 %. Some examples of text localization results are demonstrated in Fig. 13. The detected regions of text strings are marked by the blue mask. In our application, the proposed algorithm of

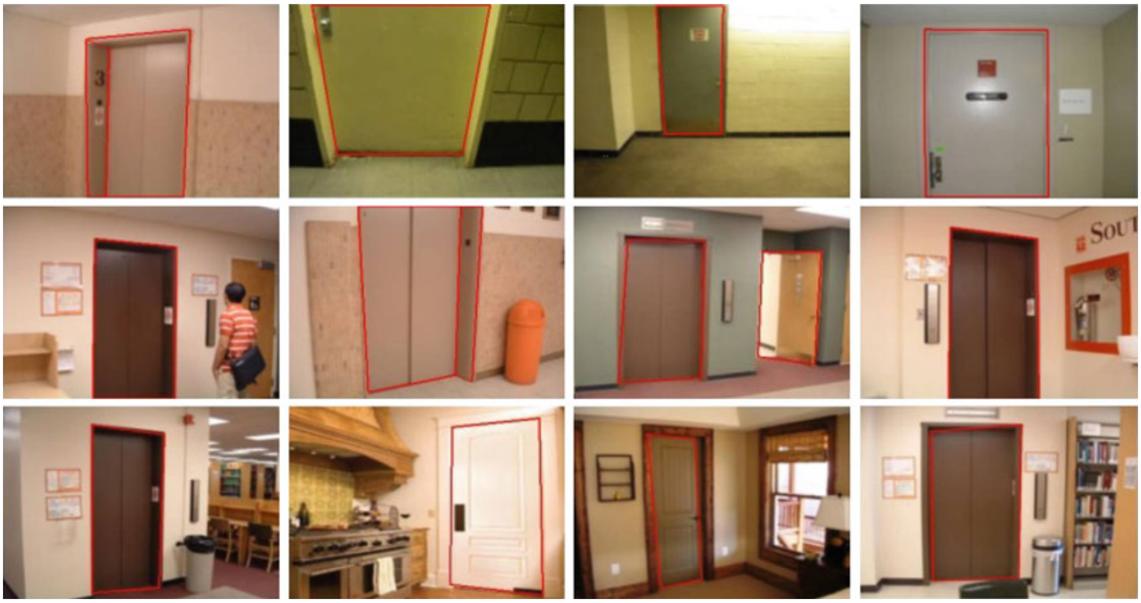


Fig. 10 Examples of successfully detected doors in different environments. The *first row* shows "Simple" examples with clear background, but includes illumination variations and occlusions. The *second row* demonstrates the "Medium" examples with more complicated backgrounds. The *third row* illustrates "Complex" examples with highly complex backgrounds

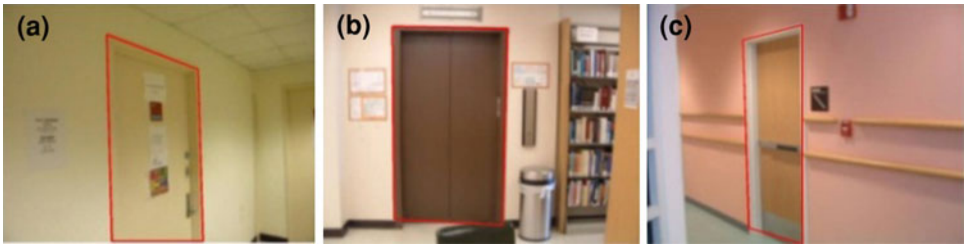


Fig. 11 Examples of door position detection: **a** *left*, **b** *front*, and **c** *right*

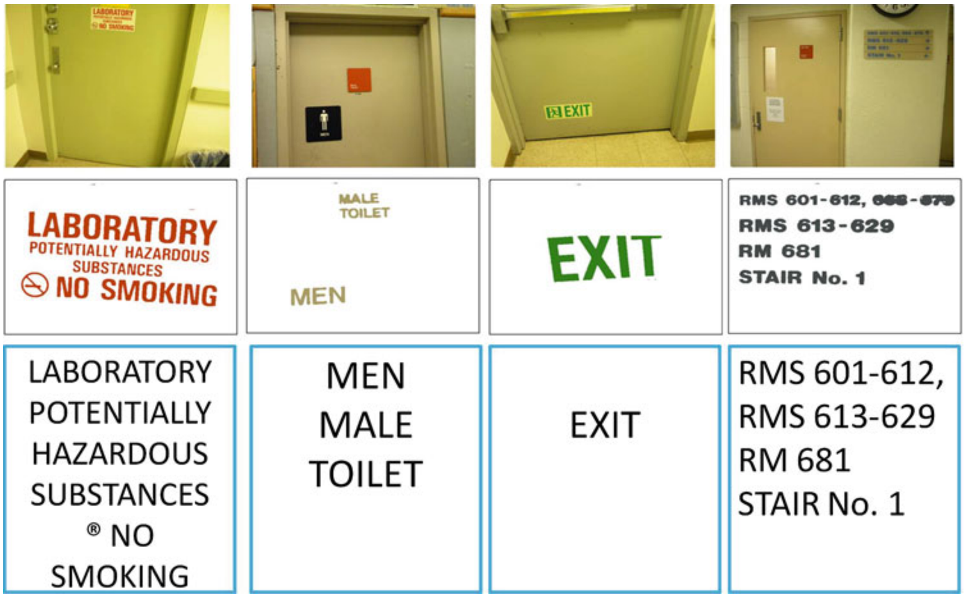


Fig. 12 *Top row* detected doors and regions containing text information. *Middle row* extracted and binarized text regions. *Bottom row* text recognition results of OCR in text regions

text region detection can achieve much better accuracy for the following two reasons: (1) relative simple background of indoor environments, and (2) only detecting text associated with the detected doors.

6.3 Discussion

Our door detection is simple and robust. It can handle both open and closed doors, because the geometric door model is based on door frame without using any appearance information of the door.

As shown in Table 2, the true-positive rate decreases and the false-positive rate increases from "Simple" to "Complex" images, since the complexity of background increases a lot. The false-positive rate of "Medium" is a little larger than that of "Complex." We believe this is due to disparity in the number of images in each category. Failure of detecting door corners appears to be the main reason for missing doors (see examples in Fig. 14a, b.)

The protrusion detection appears to be effective in lowering the false-positive rate and maintaining the true-positive rates of "Simple" and "Medium". For "Complex" images, both the rates of true positive and false positive are decreased. In images with highly complex backgrounds, some background corners adjacent to a door are detected as spurious lateral faces indicating an inset or protrusion. Such doors may be detected as protrusion objects and incorrectly eliminated. However, considering the safety of blind users, the lower false-positive rate is more desirable.

Our text extraction and localization algorithms remain subject to interference from background interference from features of color or size similar to text characters, because these pixel-based algorithms cannot filter out the information that satisfies the predefined features. Consequently, some text characters shown against complex backgrounds will inevitably lead to recognition failures or false results. For signage with fewer than three text characters in a text string, our system will fail. For example, the floor number "2" in Fig. 14b and the bathroom signage in Fig. 14c are missed.

In pilot interface studies, we have conducted a survey and collected a testing data set by ten blind subjects using a wearable camera on sunglasses (see Fig. 15). We obtain the following observations: (1) motion blur and very large occlusions happen when subjects have sudden head movements, which our method currently cannot handle; and (2) with a cane or a guiding dog, most of the blind users can

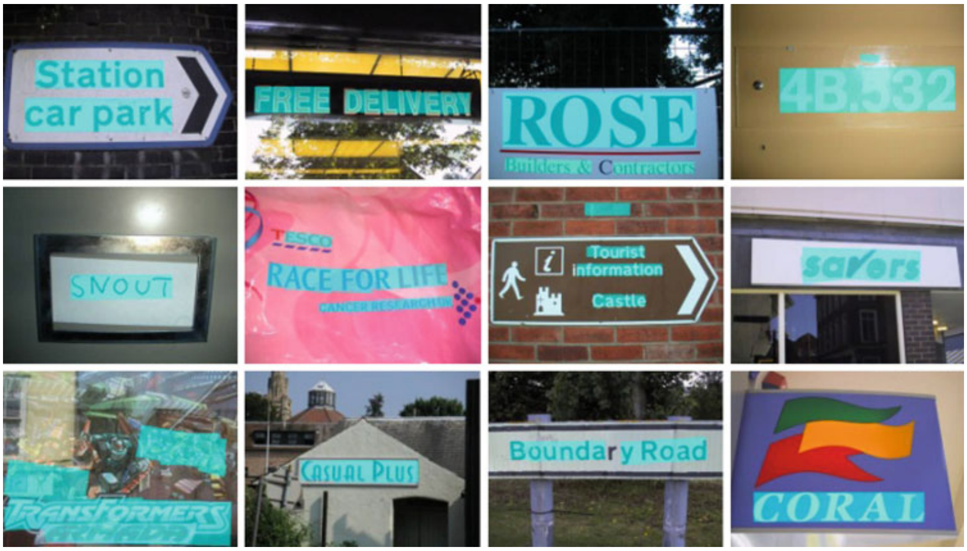


Fig. 13 Examples of text string detection on the Robust Reading Dataset. The detected regions of text strings are marked in *blue* (color figure online)

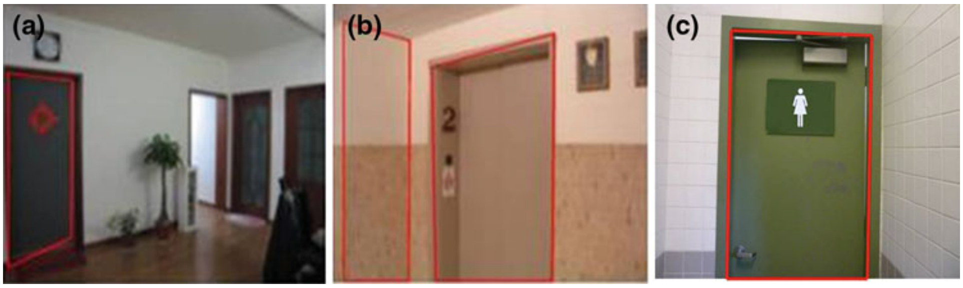


Fig. 14 Some error examples in door detection and text recognition. **a** False negative error (missing doors) for door detection. **b** False-positive error for door detection and false-negative error for text recognition of floor number "2" (only one character which does not meet the assumption). **c** False negative error for signage detection of "bathroom"

find doors without any problem. We will focus our future research on improving our text localization and recognition, signage detection and recognition, and user interface study.

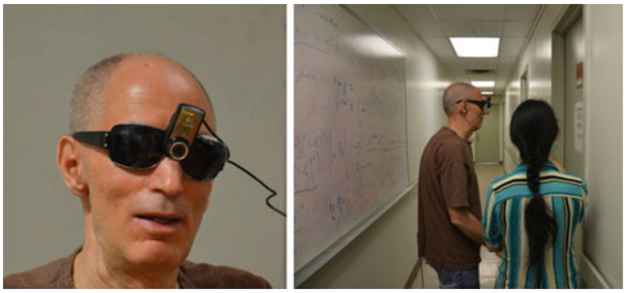


Fig. 15 Example of a prototype system includes a Logitech web camera with auto focus on sunglasses

7 Conclusions and future work

We have described some algorithms and techniques that would be important in making a prototype of computer vision-based indoor wayfinding aid to help blind persons access unfamiliar environments. Our algorithms include detection of doors and wall protrusions, and recognition text signage to differentiate doors with different functions (e.g. office from bathroom). Our novel and robust door detection algorithm is able to detect doors and discriminate other objects with rectangular shape, such as bookshelves and cabinets. Since the algorithm is based only on the general features (e.g. edges and corners) and geometric relationships, it is able to detect objects in different environments with varied colors, texture, occlusions, illumination, and viewpoints. The text signage recognition is incorporated with the detected door and further represented to blind persons in audio display. Our experiments with the door detection algorithm and text recognition in different environments have demonstrated robustness and generality of the most critical components of the proposed methods.

Our future work will focus on handling large occlusions by using more discriminative features like door knobs and other hardware, 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 comprehend the surrounding environment.

Acknowledgments This work was partially supported by NIH grant 1R21EY020990, NSF grants IIS-0957016, EFRI-1137172, and ARO grant W911NF-09-1-0565. The authors would like to thank the anonymous reviewers for their constructive comments and insightful suggestions that improved the quality of this manuscript.

References

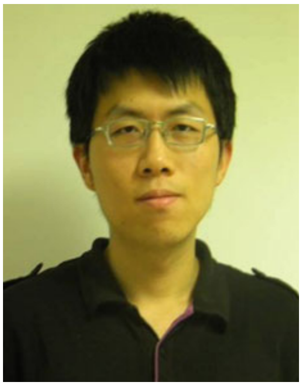
1. Arditi, A., Brabyn, J.: Signage and wayfinding. In: Silverstone, B., Lang, M.A., Rosenthal, B., Faye, E. (eds.) The Lighthouse Handbook on Visual Impairment and Vision Rehabilitation, Oxford University Press, New York (2000)
2. Anguelov, D., Koller, D., Parker, E., Thrun, S.: Detecting and modeling doors with mobile robots. In: Proceedings of the IEEE international conference on robotics and automation (2004)
3. Baker, A.: Blind Man is Found Dead in Elevator Shaft. The New York Times, City Room (2010)
4. Biederman, I.: Recognition-by-components: a theory of human image understanding. Psychol. Rev. **94** (1987)
5. Blind Sight: A camera for visually impaired people. http://accessability.blogspot.com/2008/10/blind-sight-camera-for-visually.html
6. Canny, J.: A computational approach to edge detection. IEEE Trans. Pattern Analy. Mach. Intell. PAMI **8**, 679–698 (1986)
7. Chen, X., Yuille, A.: Detecting and reading text in natural scenes, CVPR (2004)
8. Chen, Z., Birchfield, S.: Visual detection of lintel-occluded doors from a single image. IEEE Computer Society workshop on visual localization for mobile platforms (2008)
9. Chen, C., Tian, Y.: Door detection via signage context-based hierarchical compositional model. 2nd workshop on use of context in video processing (UCVP) (2010)
10. Dakopoulos, D., Bourbakis, N.G.: Wearable obstacle avoidance electronic travel aids for blind: a survey. IEEE IEEE Trans. Syst. Man Cybern. Part C Appl. Rev. **40**(1), 25–35 (2010)
11. Dinh, V., Chun, S., Cha, S., Ryu, H., Sull, S.: An efficient method for text detection in video based on stroke width similarity. Asian conference on computer vision (ACCV) (2007)
12. Divvala, S., Hoiem, D., Hays, J., Efros, A., Hebert, M.: An empirical study of context in object detection. In: Proceedings of IEEE CVPR (2009)
13. Dubey, P.: Edge based text detection for multi-purpose application. Int. Conf. Signal Process. **4** (2006)
14. Everingham, M., Thomas, B., Troscianko, T.: Wearable mobility aid for low vision using scene classification in a Markov random field model framework. Int. J. Hum. Comput. Interact. **15**(2) (2003)
15. Giudice, N., Legge, G. : Blind navigation and the role of technology. In: Helal, A.A., Mokhtari, M., Abdulrazak, B. (eds.) The engineering handbook of smart technology for aging, disability, and independence., Wiley, Hoboken (2008)
16. He, X., Yung, N.: Corner detector based on global and local curvature properties. Opt. Eng. **47**(5) (2008)
17. Hensler, J., Blaich, M., Bittel, O.: Real-time door detection based on adaboost learning algorithm. International conference on research and education in robotics, Eurobot (2009)
18. Ivanchenko, V., Coughlan J., Shen, H.: Crosswatch: a camera phone system for orienting visually impaired pedestrians at traffic intersections. 11th international conference on computers helping people with special needs (ICCHP '08) (2008)
19. Kasar, T., Kumar, J., Ramakrishnan, A.G.: Font and background color independent text binarization. Second international workshop on camera-based document analysis and recognition (2007)
20. Kim, D., Nevatia, R.: A method for recognition and localization of generic objects for indoor navigation. In: ARPA image understanding workshop (1994)
21. Kreiman, G.: Biological object recognition. Scholarpedia **3**(6), 2667. http://www.scholarpedia.org/article/Biological_object_recognition (2008)
22. Liu, C., Wang, C., Dai, R.: Text detection in images based on unsupervised classification of edge-based features. International conference on document analysis and recognition (2005)
23. Liu, Q., Jung, C., Moon, Y.: Text segmentation based on stroke filter. In: Proceedings of international conference on multimedia (2006)
24. Luo, J., Singhal, A., Zhu, W.: Natural object detection in outdoor scenes based on probabilistic spatial context models. International conference on multimedia and expo (2003)
25. Manduchi, R., Coughlan, J., Ivanchenko, V.: Search strategies of visually impaired persons using a camera phone wayfinding system. 11th international conference on computers helping people with special needs (ICCHP '08) (2008)
26. Munoz-Salinas, R., Aguirre, E., Garcia-Silvente, M., Gonzalez, A.: Door-detection using computer vision and fuzzy logic. In: Proceedings of the 6th WSEAS international conference on mathematical methods and computational techniques in electrical engineering (2004)
27. Murillo, A., Kosecka, J., Guerrero, J., Sagues, C.: Visual door detection integrating appearance and shape cues. Robot. Auton. Syst. (2008)
28. National Research Council. Electronic travel aids: new directions for research. Working group on mobility aids for the visually impaired and blind, ed. C.o. vision. National Academy Press, Washington, DC, p. 107 (1986)
29. Nikolaou, N., Papamarkos, N.: Color reduction for complex document images. Int. J. Imaging Syst. Technol. **19** (2009)
30. Oliva, A., Torralba, A.: The role of context in object recognition. Trends Cognit. Sci. **11**, 520–527 (2007)
31. Paletta, L., Greindl, C.: Context based object detection from video. In: Proceedings of international conference on computer vision systems (2003)
32. Pradeep, V., Medioni, G., Weiland, J.: Piecewise planar modeling for step detection using stereo vision. Workshop on computer vision applications for the visually impaired (2008)
33. Shen, H., Coughlan, J.: Grouping using factor graphs: an approach for finding text with a camera phone. Workshop on graph-based representations in pattern recognition (2007)
34. Shivakumara, P., Huang, W., Tan, C.: An efficient edge based technique for text detection in video frames. The eighth IAPR workshop on document analysis systems (2008)
35. Stoeter, S., Mauff, F., Papanikolopoulos, N.: Realtime door detection in cluttered environments. In: Proceedings of the 15th IEEE international symposium on intelligent control (2000)
36. Tian, Y., Yi, C., Arditi, A.: Improving computer vision-based indoor wayfinding for blind persons with context information. 12th international conference on computers helping people with special needs (ICCHP) (2010)
37. Tian, Y., Yang, X., Arditi, A.: Computer vision-based door detection for accessibility of unfamiliar environments to blind persons. 12th international conference on computers helping people with special needs (ICCHP) (2010)
38. Torralba, A.: Contextual priming for object detection. Int. J. Comput. Vision **53**(2), 169–191 (2003)
39. Tran, H., Lux, A., Nguyen, H., Boucher, A.: A novel approach for text detection in images using structural features. The 3rd international conference on advances in pattern recognition (2005)

40. Seeing with sound—the vOICe. http://www.seeingwithsound.com
41. Wan, M., Zhang, F., Cheng, H., Liu, Q.: Text localization in spam image using edge features. International conference on communications, circuits and system (2008)
42. Wong, E., Chen, M.: A new robust algorithm for video text extraction. Pattern Recognit. **36** (2003)
43. Yang, X., Tian, Y.: Robust door detection in unfamiliar environments by combining edge and corner features. 3rd workshop on computer vision applications for the visually impaired (CVAVI) (2010)
44. Zandifar, A., Duraiswami, R., Chahine, A., Davis, L.: A video based interface to textual information for the visually impaired. In: Proceedings of IEEE 4th international conference on multimodal interfaces (2002)
45. Robust Reading Dataset. http://algoval.essex.ac.uk/icdar/Datasets.html

Author Biographies

YingLi Tian received her B.S. and M.S. from TianJin University, China in 1987 and 1990 and her Ph.D. from the Chinese University of Hong Kong, Hong Kong, in 1996. After holding a faculty position at the National Laboratory of Pattern Recognition, Chinese Academy of Sciences, Beijing, she joined Carnegie Mellon University in 1998, where she was a postdoctoral fellow of the Robotics Institute. Then she worked as a research staff member in IBM T. J. Watson Research Center from 2001 to 2008. She is currently an associate professor in the Department of Electrical Engineering at the City College of New York and Department of Computer Science at the Graduate Center, the City University of New York. Her current research focuses on a wide range of computer vision problems from motion detection and analysis, assistive technology, to human identification, facial expression analysis, and video surveillance. She is a senior member of IEEE.

Xiaodong Yang received his B.S. degree from Huazhong University of Science and Technology, Wuhan, China in 2009. He is currently pursuing his Ph.D. degree in electrical engineering at the City College, City University of New York. His current research focuses on recognition of objects, texture, and scene categories. His research interests include image processing, object detection, and developing effective image representations for recognition.



Chucai Yi received his B.S. and M.S. degrees in Department of Electronic and Information Engineering from Huazhong University of Science and Technology, Wuhan, China, in 2007 and 2009, respectively. From 2009, he has been a Ph.D. graduate student in computer science at the Graduate Center, the City University of New York, New York, NY, USA. His research focuses on text detection and recognition in natural scene images. His research interests include object recognition, image processing, and machine learning

Aries Arditi, Ph.D. is Principal Scientist at Visibility Metrics LLC, and President of the Mars Perceptrix Corporation. Formerly a senior fellow in vision science and Vice President for vision science at Lighthouse International and a research staff member at IBM TJ Watson Research Center, he has authored over 100 scientific publications. His research interests include visual accessibility of the built environment, computers and internet, text legibility, color and low vision, reading and low vision, assessment of vision function, wayfinding, binocular vision, functional perimetry, and prosthetic vision. His contributions include seminal work in the analysis of visual field defects in binocular visual space (volume perimetry), the first published studies of the independent impact of font parameters on legibility, and numerous studies on the psychophysics of reading and of functional assessment of vision. He is the creator of the Mars Contrast Sensitivity Tests used throughout the world; and of LowBrowse^{TM}, a unique, open source add-on to the Mozilla Firefox browser that makes it easier for visually impaired people to access the web. Dr. Arditi has served as President of the International Society of Low Vision Research and Rehabilitation (two terms), as Editor-in-Chief of the Informa Healthcare journal Visual Impairment Research (10 years), and on numerous government panels and committees. He is a Diplomate (low vision research) and Fellow of the American Academy of Optometry and a Fellow of the American Psychological Society.