Recognizing Elevator Buttons and Labels for Blind Navigation

Jingya Liu and Yingli Tian, *Senior Member*, IEEE
Dept. of Electrical Engineering
The City College of New York
NY 10031, USA
jliu008@citymail.cuny.edu, ytian@ccny.cuny.edu

Abstract- **In this paper, we propose a method to detect elevator buttons and recognize their labels from images for blind navigation. First, a pixel-level mask of elevator buttons is segmented based on fully convolutional networks. Then a fast scene text detector is applied to recognize the text labels in the image as well as to extract their spatial vectors. Finally, all the detected buttons and their associated labels are paired by combining the button mask and spatial vectors of labels based on their location distribution. To evaluate the proposed method, we collect an elevator button dataset that contains 1,000 images with buttons captured from both inside and outside of elevators and annotate the locations and labels of all buttons. Preliminary results demonstrate the robustness and effectiveness of the proposed method for elevator button detection and associated label recognition.**

I. INTRODUCTION

Independent travel presents significant challenges for blind or vision-impaired person. Many research efforts have been conducted to help blind or visually impaired people with daily activities. One of the research areas is indoor navigation and wayfinding [1-2]. Tian et al. developed an assistive indoor navigation system by detecting doors and elevators and recognizing the corresponding text descriptions. An RGBD camera and feedback from obstacles contribute to a wearable system for safely guiding blind people in a walkable space [3]. Thanks to the rapid increasing computer vision techniques, robust object recognition methods make it possible to develop more reliable assistance systems. Aerial obstacle detection using a 3D smartphone achieved a real-time detection of the overhead objects (e.g. branches or awnings) which cannot be detected by a white cane or a guiding dog for visual impaired people [4]. Recently, robot guides were designed with navigation variety functions for blind people in indoor environments [5].

For indoor navigation and wayfinding, elevators are the most common tool to access multiple floors. Some standards have been implemented for elevator design to help blind users. For example, braille descriptions for tactile reading are mandatorily located nearby texts or symbols [12]. Although there are braille labels on elevator button panels to assist blind users, it is still very challenging for them to locate the elevator buttons of the floors they want to go. Therefore, this paper focuses on developing a method to accurately and efficiently detect elevator buttons and recognize the associated text labels from images.

Detecting elevator buttons from images is a challenging task due to the following reasons as shown in Figure 1: 1) high similarity to the background. For example, the buttons are often made by same material as the background. 2) Dark lighting inside elevators. 3) High variety of button designs and layouts. Our method recognizes both call buttons outside elevators (up or down) and buttons inside elevators in a variety of formats regardless of different background textures and colors, variety button designs and layouts, and different capture viewpoints. In addition, our method provides the spatial locations of elevator buttons and their associated labels which are very important to guide blind users.

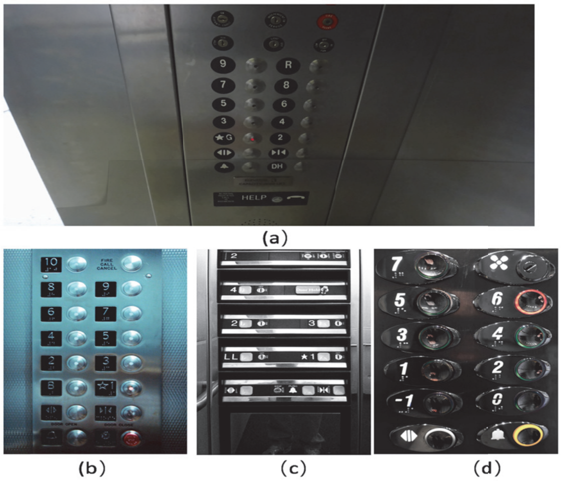

Fig. 1. Examples in our collected elevator button dataset.

The rapidly developed techniques of image-based object detection, recognition, and semantic segmentation can be applied to our task. In order to detect locations of elevator buttons, fully convolutional networks (FCNs) have been adopted to achieve accurate pixel-level segmentation results without losing the important object spatial information [6]. The state-of-the-art semantic segmentation methods can segment and classify different objects. Furthermore, instance-aware semantic image segmentation achieves great performance on differentiating each object from same category. Multi-task Network Cascades are employed to conduct individual object detection, segmentation, and object categorization, shared by same feature map predicting an accurate and efficient instance segmentation [7]. Natural language description plays an important role on identifying different objects. By combining with visual and linguistic information, object semantic segmentation is capable to process complicated queries. For example, with a query of '*a man on the right*' and an image, it returns a segmentation result for a man on the right side [8]. A prior language description is required for this instance segmentation task. In our application, the prior language description can be the speech commands from the blind user such as "button for the 8th floor".

Inspired by the previous work [6-9], we develop a framework for button detection and label text recognition. This paper combines the semantic segmentation mask of elevator buttons with the spatial information for the corresponding text descriptions adjacent to buttons.

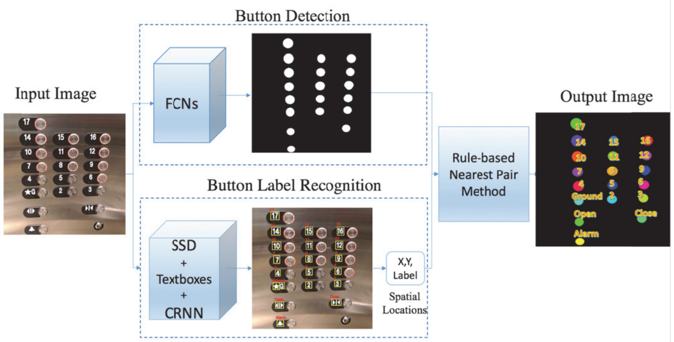

Fig. 2. The flowchart of the proposed framework for button detection and label recognition. Fully convolutional networks (FCNs) generate a pixel-level mask for buttons. A set of spatial vectors of labels (*X, Y, Label*) is computed by a single-shot detector (SSD) and convolutional recurrent neural network (CRNN). A rule-based nearest pair method combines button mask with spatial vectors of labels to match the buttons and their associated labels.

II. Elevator Button Detection and Label Recognition

As shown in Fig. 2, our method takes an image as input and outputs the locations of buttons represented by an instance-aware semantic segmentation image with marked labels. Our method consists of three main components: 1) Elevator button detection. This component estimates a semantic segmentation mask for elevator buttons at pixel-level using Fully Convolutional Networks (FCNs). 2) Elevator button label recognition. The text descriptions and spatial locations of button labels are recognized by a single-shot detector (SSD). 3) Elevator button identification. By applying a rule-based method to search the nearest label around the button, each button is then associated with its corresponding label. More details of each component are described in the following sections.

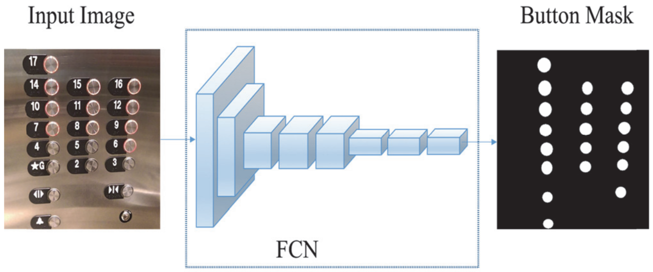

Fig. 3. The flowchart for FCNs-based elevator button detection.

A. Elevator Button Detection.

In order to recognize elevator buttons, the first step is to detect where elevator buttons are in the input image. Compared to region level-based object detection and recognition, pixel level-based semantic segmentation can provide more accurate location and shape information of objects to benefit blind navigation.

Fully Convolutional Networks (FCNs) are widely employed in image segmentation and object classification [13]. Similar to [13], we train a network with fully convolutional layers. As shown in Fig. 3, it takes an image as the input and outputs the pixel level segmentation mask which classifies the elevator buttons from background. The elevator button mask contains pixel-level segmentation regions of elevator buttons and reserves their location and shape information.

This paper implements FCN32s, which keep the first five convolution layers from VGG-16 architecture [13] and then followed by three convolutional layers, where the 32s indicates upsamples stride number to recover the size of original image due to the downsampling made by the pooling layers. In this network, the input of each layer contains a three-dimensional matrix of image height, width, and channel feature, which are varied from layer to layer. Therefore, FCN retains the spatial information of image. The pre-trained VGG-16 model on ImageNet is fine-tuned on our elevator button dataset.

B. Elevator Button Label Recognition.

To recognize button labels, we conduct a text detection method. As shown in Fig. 4, we first detect the button related text information of numbers, letters, symbols of the location information of labels, and then recognize them in order to combine with the mask generated from elevator button detection.

A single-shot detector (SSD) based network [10] is integrated with non-maximum suppression to generate a set of text candidates at region level including letters and numbers. The locations and sizes of the detected text regions are represented by bounding boxes. This network is a depth single neural network model adopted for target detection and recognition. It uses a multi-scale characteristic feature detection network. In order to improve the text detection accuracy, a new Textbox layer, is added to SSD network, which can deal with multiple text regions with arbitrary sizes. The outputs of the Textbox layer are refined text bounding box candidates. The loss function for SSD is defined as the sum of localization loss (L_{loc}) and confidence loss (L_{conf}):

$$\mathrm{L(x,c,b,g)} = \frac{1}{N}\left(L_{conf}(x,c) + \alpha L_{loc}(x,b,g)\right),$$

where N is the number of matched default boxes with pre-defined value for box detection, x is the classification value, b is the predicted bounding box, g represents the ground truth box, α is the weight term which is set to 1, and c indicates class confidences.

Then, the refined text candidates are fed a state-of-the-art text recognizer named convolutional recurrent neural network (CRNN) to recognize the recognition of numbers, letters and symbols [11]. CRNN is a trainable network using jointly combination of Residual Neural Network and Convolutional Neural Network with arbitrary length inputs (arbitrary width and length of text candidate regions). The advantage of this method is that it preserves an outstanding performance while using a model with less layers and parameters.

The outputs of CRNN can be represented as a vector contains the spatial information for the text regions and the recognized text descriptions.

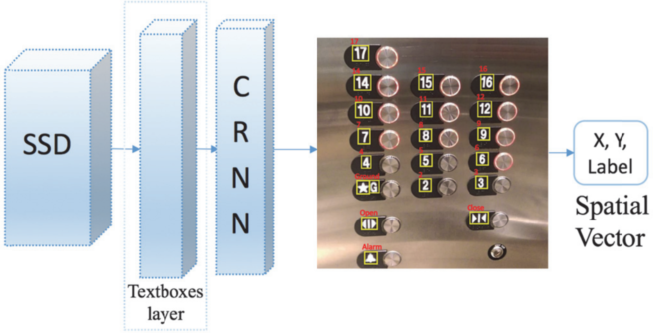

Fig. 4. The flowchart for recognizing text labels of elevator buttons.

C. Elevator Button Identification

After obtaining elevator button mask and recognizing text labels, we need to identify each elevator button with the corresponding label. In general, the button label is located in the closest distance from its corresponding button compared to surrounding buttons. Therefore, we develop a straightforward rule-based pair method to match the buttons and labels by combining the pixel-based elevation button mask and the spatial vector of the recognized text labels. For each detected button, we calculate the Euclidian distance between the centers of the button and the recognized labels. The label with the closest distance is identified as the corresponding label of the button.

III. Experimental Results and Discussions

Dataset: There is no available public dataset for image-based elevator button detection and recognition. In order to evaluate the proposed method, we collect a new dataset contains 1,000 images for both elevator call buttons (outside the elevator) and control panels (inside the elevator). We annotate each image including pixel-level *button* regions and object-level text descriptions. The images in the dataset contain elevator buttons ranging from 1 to 70 with variety of shapes, textures, and layouts. In our experiments, 80% images from the dataset are randomly selected for training and validation and the remaining 20% images are used for testing.

Preprocessing: To handle the lighting changes, we first apply a preprocessing on the input image. Due to insufficient light in elevators, buttons and labels often appear too dark to be detected or recognized. Therefore, a histogram equalization process is first applied to reduce the light variations.

A. Results of Elevator Button Detection.

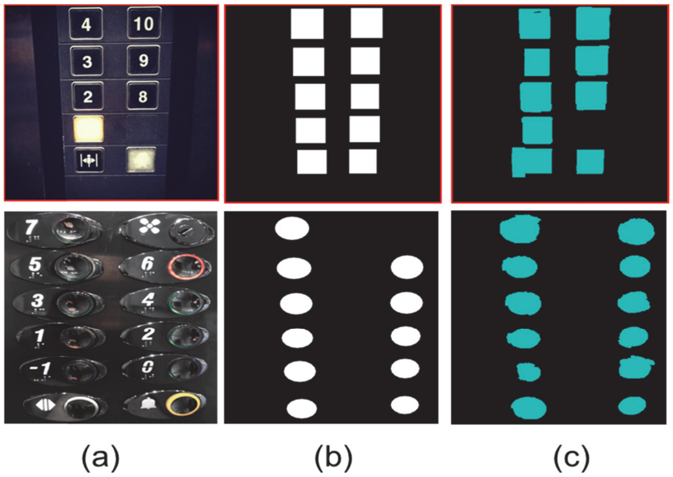

Fig. 5 Example results for Elevator Button Detection. (a) Input image. (b) Ground-truth (c) Our detection result.

We fine-tune the pretrained FCN model on our collected dataset which contains two categories: *button* and *background*. The input takes arbitrary sizes of images. For the first convolution layer, we pad 100 pixels around the image in order to guarantee the sixth layer convolution image is not less than the size of 192 × 192. The offset value is set to 19 to guarantee upsampling the images to the original sizes. Examples of elevator button detection results are demonstrated in Fig. 5. It shows that the button regions are clearly separated. However, due to the very similar shape of buttons and labels, some labels are wrongly detected as buttons. This issue can be solved by combining with the label spatial vectors. The accuracy is calculated by intersection over union of detected button and ground-truth and achieves 73.2% correction rate.

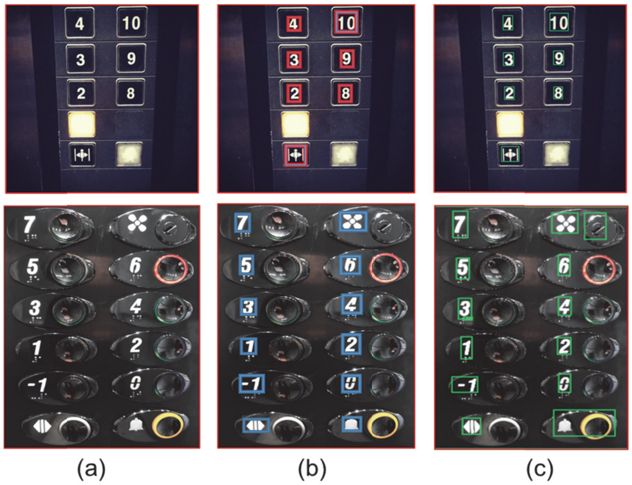

Fig. 6 Examples of Elevator Button Label Detection and Recognition. (a) Input image. (b) Ground-truth (c) Our result.

B. Results of Elevator Button Label Recognition.

For elevator label recognition, the network resizes the input image to 300 × 300 pixels. VGG-16 is employed as the pre-trained network with retaining the top 5 layer structures. From the 6th layer, the SSD block is joined. For the 6th and 7th layers, 3 × 3, 1 × 1 convolution kernel sizes are used respectively. From the 8th to 11th layers, each layer applies a convolution with 1× 1 kernel size overlay 3 × 3. Text boxes are used in multidimensional text detection for different lengths and the sizes of texts in an images, and the sizes are fixed to {1 ×1, 2 ×2, 3 ×3, 5 ×5, 7 ×7, 10 × 10}. CRNN testing scores is defined as $s = \max_{w \in \mathcal{W}} p(w|I)$, where I indicates the input image, w is the character sequence, and $\mathcal{W}$ is the given lexicon. Some examples of the text detection results are shown in Fig. 6, most of numbers, letters, symbols of the button are detected. The accuracy is calculated by comparing the percentage of labels are correctly detected and recognized, which is 71.9% in our experiment.

C. Results of Elevator Button Identification

By combining the elevator button mask and the spatial vector of the recognized labels, each button and its corresponding label is identified based on the shortest distance rule. Some examples of the identification results are shown in Fig. 7, while buttons are displayed in different colors indicate the correspondence to different labels. We obtain an accuracy rate at 70.31% by comparing our identification results with the ground-truth.

Fig. 8 shows more examples for the 3 main components in our network. First, for elevator button detection (Fig. 8 (a)-(c)), the results demonstrate the ability to accurately segment the regions of buttons. However, for the situation where the label panel has the similarity layout with buttons, some labels are identified as buttons. Second, most of the labels are successfully detected on the label panels and buttons as shown in Fig. 8 (d)-(e). The results demonstrate that regions of label candidates may overlap with other regions. The result is acceptable due the delamination from button identification. Fig. 8 (f)-(g) shows the ground-truth and the final results of button and label pairs. Most buttons are successfully identified. However, it can be observed that for the second row and the last row, due to the similar layout of buttons and labels, a region of the label panels is also considered as button. These cases are acceptable to our system while the blind user can locate the button by touching the corresponding label panel.

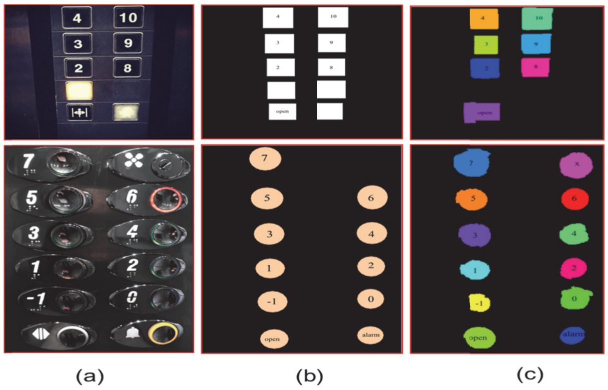

Fig. 7 Examples of Elevator Button Label Identification by pairing buttons and the corresponding labels. (a) Input image. (b) Ground-truth. (c) Our result.

IV. Conclusions

We have presented a new method for assistive blind navigation by combining the object semantic segmentation with text recognition to detect and recognize elevator buttons and labels using a cascade network. The preliminary results are very promising. We achieve 70.31% accuracy on detecting the correlation between button and label. The proposed method can further be extended to segment objects with associated text descriptions for many applications. Our future work will focus on implementing the proposed method on a mobile device, developing a user-friendly interface, providing effective audio feedback about the location and label of the queried elevator button, and evaluating the developed system and interface by blind users.

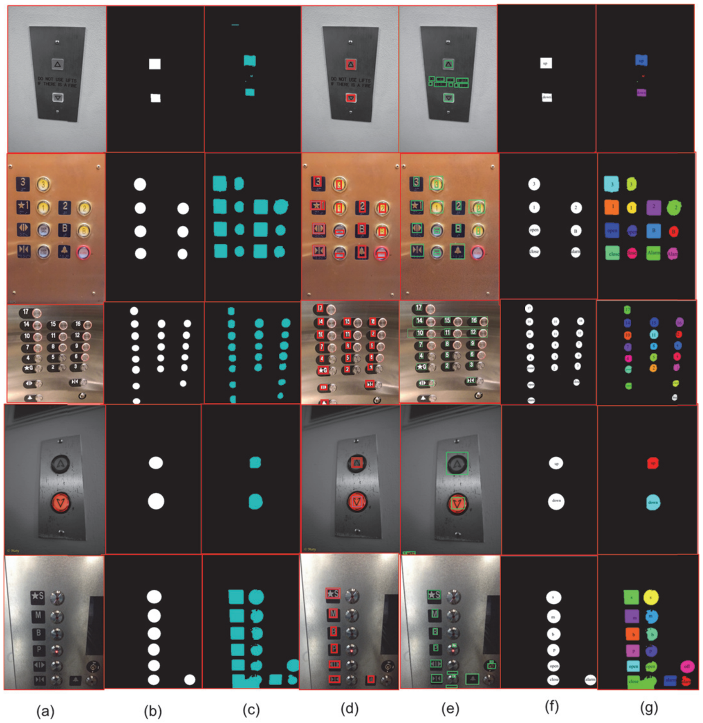

Fig. 8 Elevator Button Label Recognition. (a) Input image. (b) Ground-truth of button detection. (c) The result of button recognition. (d) Ground-truth of label. (e) The result of label recognition. (f) Ground-truth of button identification. (g) The result of button identification.

Acknowledgment

This work was supported in part by NSF grants EFRI-1137172, IIP-1343402, and IIS-1400802.

References

[1] Y. Tian, X. Yang, C. Yi, and A. Arditi, "Toward a computer vision-based wayfinding aid for blind persons to access unfamiliar indoor environments." Machine Vision and Applications, Vol.24, No.3 (2013): 521-535.

[2] X. Rong, B. Li, J. P. Mũno, J. Xiao, A. Arditi and Y. Tian, "Guided Text Spotting for Assistive Blind Navigation in Unfamiliar Indoor Environments", 12th International Symposium on Visual Computing (ISVC), 2016.

[3] S. Tobias, M. Lauer, M. Schwaab, M. Romanovas, S. Böhm, and T. Jürgensohn, "A camera-based mobility aid for visually impaired people." KI-Künstliche Intelligenz, Vol.30, No. 1 (2016): 29-36.

[4] J.M. Sáez, F. Escolano, and M.A. Lozano, "Aerial obstacle detection with 3-D mobile devices." IEEE Journal of Biomedical and Health Informatics, Vol.19, No.1 (2015): 74-80.

[5] C. Feng, S. Azenkot, and M. Cakmak, "Designing a robot guide for blind people in indoor environments." ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts, 2015.

[6] J. Long, E Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation." CVPR, 2015.

[7] J. Dai, K. He, and J. Sun, "Instance-aware semantic segmentation via multi-task network cascades." CVPR, 2016.

[8] R. Hu, M. Rohrbach, and T. Darrell. "Segmentation from natural language expressions." ECCV, 2016.

[9] M. Liao, B. Shi, X. Bai, X. Wang, and W. Liu, "TextBoxes: A Fast Text Detector with a Single Deep Neural Network." arXiv preprint arXiv:1611.06779 (2016).

[10] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Fu, and A.C. Berg, "SSD: Single shot multibox detector." ECCV, 2016.

[11] B. Shi, X. Bai, and C. Yao, "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition." IEEE Transactions on Pattern Analysis and Machine Intelligence (2016).

[12] ADA Standards for Accessible Design, 2010 Standards, (2010).

[13] K. Simonyan, and A. Zisserman, “Very deep convolutional networks for large-scale image recognition.” arXiv preprint arXiv:1409.1556, (2014).