International Journal of Computer Applications (0975 - 8887)
Volume * - No.*, —— 2020

Recognizing Elevator Buttons and Labels for Blind Navigation

Jingya Liu
The City College of New York, NY 10031 USA
jliu1@ccny.cuny.edu

Yingli Tian*
The City College of New York, NY 10031 USA
ytian@ccny.cuny.edu

ABSTRACT

In this paper, a cascade framework is proposed 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 deep neural 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. The cascade framework is conducive to multitask but the accuracy may decrease task by task. To avoid the limitation of the intermediate task, a new schema is further introduced by pairing buttons with their labels to consider the region of button and label as a whole. First, the regions of button-label pairs are detected and then the label for each pair is recognized. To evaluate the proposed method, an elevator button detection dataset is collected including 1,000 images containing buttons captured from both inside and outside of elevators with annotations of button locations and labels and 500 images are captured in elevators but without button buttons which are used for negative images in the experiments. Preliminary results demonstrate the robustness and effectiveness of the proposed method for elevator button detection and associated label recognition.

General Terms

Computer Vision, Pattern Recognition

Keywords

Object detection, Semantic segmentation, Computer vision, Deep learning

1. INTRODUCTION

Independent travel presents significant challenges for the blind or vision-impaired person. Many research efforts have been conducted to help blind or visually impaired people with daily activities such as grocery shopping [1], indoor navigation and wayfinding [2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12], and outdoor navigation [13]. Tian et al. [1] 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 [14, 15]. RGBD cameras are also employed for staircase detections to assist blind users [8, 9]. The RFID technology is used to determine the location for blind people [16, 17] and spatial knowledge is constructed by auditory virtual reality [18].

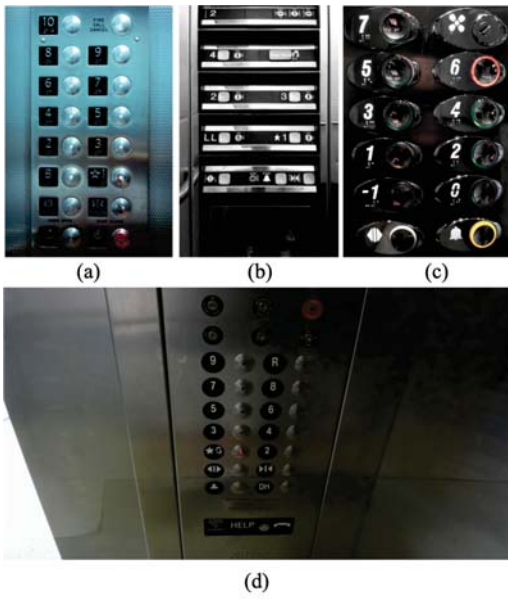

Fig. 1. Examples in our collected elevator button dataset.

Thanks to the rapidly 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 [20, 21]. Recently, robot guides or smart canes were designed with navigation variety functions for blind people in indoor environments [12, 22].

For indoor navigation and wayfinding, elevators are the most common tool to access multiple floors. Some standards have been

*Corresponding author.

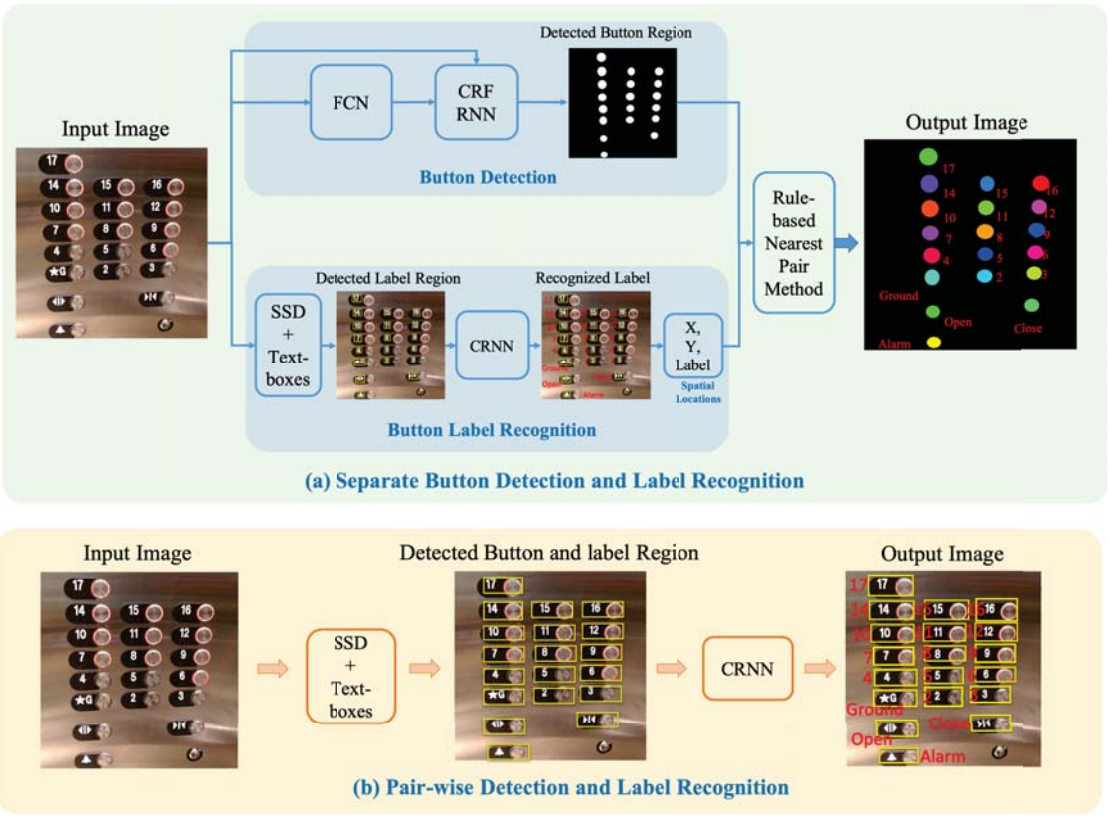

Fig. 2. The two frameworks for the detection and recognition of elevator buttons and labels. (a) The flowchart of the proposed framework for detecting and recognizing buttons and labels separately. Conditional random fields as recurrent neural networks (CRFasRNN) generate a pixel-level mask for buttons. A set of spatial vectors of labels (X, Y, Label) is computed by a single-shot detector (SSD) [19] 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. (b) The flowchart of the paired button and label detection applying SSD and label recognition using CRNN.

Fig. 3. Examples in our dataset without elevator button panel (refer as backgrounds).

implemented for elevator design to help blind users. For example, braille descriptions for tactile reading are mandatorily located nearby texts or symbols [23]. Although there are braille labels on elevator button panels to assist blind users, it is still very inconvenient for them to locate the elevator buttons of the floors they want to go. Therefore, a method is developed to accurately and efficiently detect elevator buttons and recognize the associated text labels from images.

An earlier version of this paper with preliminary results can be found in [24]. Compared to our previous work, there are three major extensions that merit being highlighted: 1) the previous work separately detects elevator buttons and their labels. A new united method proposed in this paper can handle both the elevator button and the associated label as a pair which can achieve higher accuracy; 2) the previous work was not able to handle negative images (i.e. images containing no elevator buttons). In this paper, the method is extended to detect if there are buttons in the camera view. This process significantly reduces false recognition rate and computation cost; and 3) we enlarge the evaluation dataset and add more experimental results to demonstrate the efficiency and robustness of the proposed work.

Detecting elevator buttons from images is a challenging task due to the following reasons as shown in Fig. 1: 1) Dark lighting inside elevators. 2) The high variety of button designs and layouts. 3) high similarity to the background. For example, the buttons are often made by the same material as the background. The proposed method recognizes both call buttons outside elevators (up or down) and buttons inside elevators in a variety of formats regardless of dif-

ferent background textures and colors, variety button designs and layouts, and different capture viewpoints. In addition, the proposed method provides the spatial locations of elevator buttons and their associated labels which are very important to guide blind users.

The rapidly developed techniques of image-based object detection, recognition, and semantic segmentation can be applied to the elevator recognition and detection 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 [25]. 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 the same category. Multi-task Network Cascades are employed to conduct individual object detection, segmentation, and object categorization, shared by the same feature map predicting an accurate and efficient instance segmentation [26].

Natural language description plays an important role in 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 [27]. 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 [25, 27, 28], a framework is developed 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.

2. ELEVATOR BUTTON DETECTION AND LABEL RECOGNITION

As shown in Fig. 2(a), the proposed method takes an image as input and outputs the locations of buttons represented by an instance-aware semantic segmentation image with marked labels. The framework consists of three main components: 1) Elevator button detection. This component estimates a semantic segmentation mask for elevator buttons at pixel-level using conditional random fields as recurrent neural networks (CRFasRNN). 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. In Fig. 2(b), a novel method is further proposed that first takes input image using SSD to detect pair-wised button and label. Then CRNN is applied for symbol recognition

2.1 Elevator Button Preselection

Due to the variety of images captured by cameras, the system may detect irrelevant scenes of similar objectives or categories, which leads to a high false-positive rate, as well as the waste of computation time and memory. The elevator button preselection process is used to distinguish the certain area of the elevator panel avoiding redundant detection on every category of scenes. To solve the problem, a dataset is collected which contains two main categories of images with or without button panels and annotates the elevator images with buttons as the 'button' category and without button panel as 'background' category. The examples of elevator panels without buttons and elevator background are shown in Fig. 3. A finetune a convolutional neural network using VGG16 [29] which contains 16-layer convolutional networks is used for classification and trained on the dataset contains only two categories: elevator images with button panel and without button panel.

2.2 Separate Button Detection and Recognition

2.2.1 Elevator Button Detection. In order to recognize elevator buttons, this 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 a 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 [29]. Follow the work of [29], FCNs take an image as the input and output the pixel level segmentation mask to classify the elevator buttons from the background. The elevator button mask contains pixel-level segmentation regions of elevator buttons and reserves their location and shape information.

While the button appeared in images with very clear boundaries, FCNs perform well in detection tasks for most of the cases. However, some of the buttons and backgrounds have a blurred boundary due to the illumination problem or the low-resolution ratio of images. It requires the network to be more precise on detecting the edge of button, keeping both appearance consistency and spatial consistency. In this paper, another exiting network is adopted named Conditional Random Fields as Recurrent Neural Networks (CRFasRNN) [30]. CRFasRNNs combine DenseCRF [31] with FCN iterations to keep feature representations. CRF can extract weak and coarse forecasting to produce sharp boundaries and fine-grained (precision fine lines). It utilizes the pixel label marking probability prediction problem, which contains the hypothesis of similar pixels between tag consistencies. CRFasRNN uses the mean-field approximation process of CRFs as Recurrent neural network (RNN) iteration and embedding CRFasRNN to CNNs to get forward propagation and backpropagation training model. Fig. 2(a) shows the detail of the button detection network.

2.2.2 Elevator Button Label Recognition. To recognize button labels, a text detection method is conducted. As shown in the button label recognition network of Fig. 2(a), firstly, the button related text information of a number is detected, including the 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 [19] is integrated with non-maximum suppression to generate a set of text candidates at the regional 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}):

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

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 groundtruth box, is the weight term which is set to 1, and c indicates class confidences.

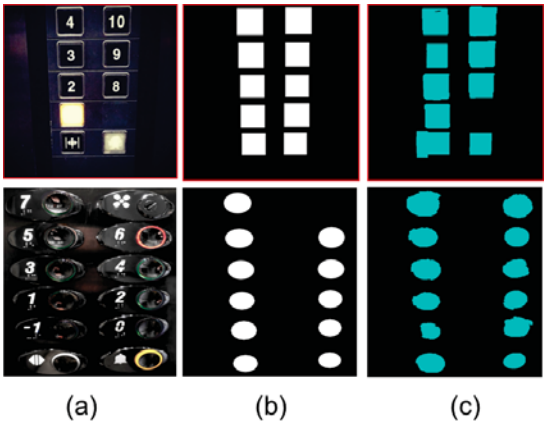

Fig. 4. Example results for Elevator Button Detection. (a) Input images. (b) Ground-truths. (c) Our results.

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 [32]. CRNN is a trainable network using jointly combination of Recurrent 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 fewer 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.

2.2.3 Elevator Button Identification. After obtaining the elevator button mask and recognizing text labels, identifying each elevator button with the corresponding label is needed. In general, the button label is located in the closest distance from its corresponding button compared to the surrounding buttons. Therefore, a straightforward rule-based pair method is developed 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, the Euclidian distance is calculated 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.

2.3 Pair-wise Button Detection and Recognition

Although the proposed model described in previous sections achieves a promising accuracy with cascades network using three main components, the accuracy eliminates due to the accumulated loss of features through each component step. To overcome this issue, button and label is considered as a whole area that provides strong button and label features and additionally and avoids the false positive detection of button shape objects. The examples of paired buttons and labels are shown in Fig. 5.

As shown in Fig. 2(b), SSD network is employed to detect and generate text candidate regions for the pair-wise button and label

Fig. 5. Examples of the results for pair-wise button and label detection.

for recognition. For example, on the upper left corner of the input image shown in Fig. 2(b), the round button and black panel with number "17" on top are detected as one object. CRNN is then applied to recognize the semantic meaning, which refers to digital numbers, or symbols in these regions of each labeled panel in the candidate region.

3. 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, a new dataset is collected which contains totally 1,500 images for 1,000 images with both elevator call buttons (outside the elevator) and control panels (inside the elevator), and 500 background images contain the elevator scene without button panel. The annotation for each image includes both pixel-level button regions and object-level text descriptions. The number of elevator buttons in the images of the dataset ranges from 1 to 70 with a variety of shapes, textures, and layouts. The 80% images from the dataset are randomly selected for training and validation, and the remaining 20% images are used for testing.

3.1 Results of Elevator Button Preselection

To handle the lighting changes, a histogram equalization process is applied to the input images. For the preselection model using VGG16, the model achieves 92.32% accuracy on successful classification comparing predicted class with ground-truth. More details will be discussed in the following sections.

3.2 Results of Separate Elevator Button Detection and Recognition

3.2.1 Results of Elevator Button Detection. The pre-trained CRFasRNN model is finetuned on our collected dataset which contains two categories: button and background region in an image. The input takes arbitrary sizes of images. Examples of elevator button detection results are demonstrated in Fig. 4. 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 the union of the detected button and ground-truth and achieves 74.2% accuracy by calculating the percentage of the mean Intersection over Union (IoU) for the predicted button region and Ground-truth larger than 0.5.

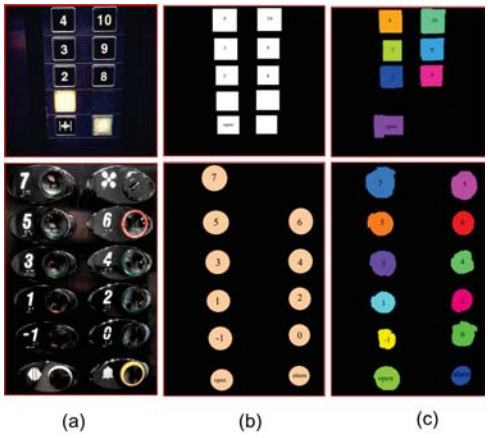

Fig. 6. Examples of Elevator Button Label Identification by Pairing Buttons and the Corresponding Labels. (a) Input images. (b) Ground-truths. (c) Our results.

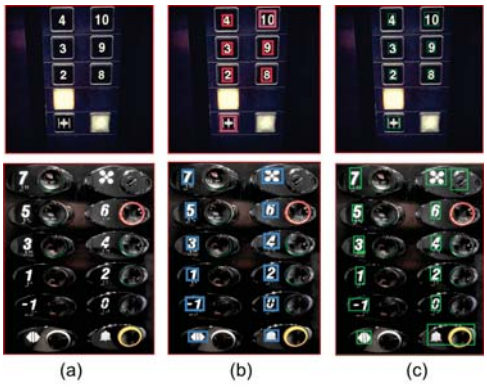

Fig. 7. Examples of Elevator Button Label Detection and Recognition. (a) Input image. (b) Ground-truths. (c) Our results.

3.2.2 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 6^{th} layer, the SSD block is joined. For the 6th and 7^{th} layers, 3×3, 1×1 convolution kernel sizes are used respectively. From the 8_{th} to 11_{th} 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 image, and the sizes are fixed to $1 \times 1, 2 \times 2, 3 \times 3, 5 \times 5, 7 \times 7, 10 \times 10$. CRNN testing scores are defined as where I indicate the input image, ω is the character sequence, and is the given lexicon. Some examples of the text detection results are shown in Fig. 7, most of the numbers, letters, symbols of the button are detected. The accuracy is calculated by comparing the percentage of labels that are correctly detected and recognized, which is 71.9% in the experiment. The reasons for unsuccessful recognized labels are mainly caused by: 1) the images contain low quality, blurry or wear labels; 2) the images with low illumination have unsharp edge pixels for buttons or labels, and 3) several symbols may appear with limited number in the training dataset and feature map accordingly are lost during the training process which cannot be recognized by the model.

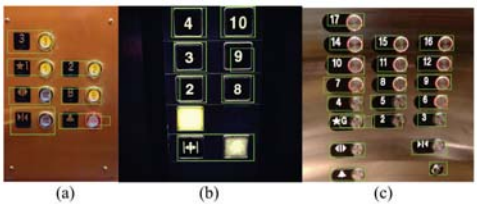

Fig. 8. Results of pair-wise button and label detection.

3.2.3 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. 6, while buttons are displayed in different colors indicate the correspondence to different labels. the proposed framework obtains an accuracy rate at 73% by comparing the identification results with the groundtruths.

Fig. 9 shows more results from the 3 main components in the proposed network. First, for elevator button detection (Fig. 9(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. 9(d)-(e). The results demonstrate that regions of label candidates may overlap with other regions. The results are acceptable due to the delamination from button identification.

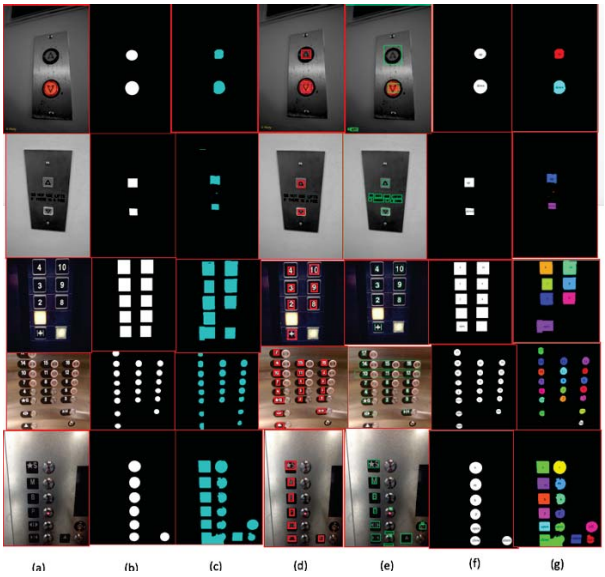

Fig. 9. Elevator Button Label Recognition. (a) Input images. (b) Ground-truths of button detection. (c) The results of button recognition. (d) Ground-truths of label. (e) The results of label recognition. (f) Ground-truths of button identification. (g) The results of button identification.

Table 1. Accuracies and false positive rates calculated for pair-wise button and label detection at different mean IoUs.

IoU	Accuracy	False positive
0.5	75.3%	16.7%
0.45	79.1%	18.2%
0.4	79.6%	22.3%

3.3 Results of Pair-wise Button Detection and Recognition

As shown in Fig. 8(a)-(c), the proposed model successfully detects and recognizes the pair-wise button and label. We observe that for several cases (e. g. the lower right corner key button in Fig. 8(c)), the model detects the round-shaped key as button-label pairs. The reason for the false-positive detection is that the round-shaped key-hole looks like a label. However, the situation can be neglected as blind users can easily distinguish real and fake buttons.

As shown in Table 1, the baseline model for train the pair-wise button and label using 0.5 mean Intersection over Union (IoU) boosting the accuracy up to 75.3% which raising around 5% compared to three component-based models with 16.7% false positive. Similarly, the IoU is set to 0.45, the accuracy increases to 79.1% with 18.2% false positive. The accuracy reaches to 79.6% while the IoU equals to 0.4 with 22.3% false positive. With the relatively stable false positive number and the increasing accuracies with the decreasing IoU, the pair-wise button and label with 0.4 IoU achieve the best performance.

4. DISCUSSIONS

Table 2 shows the results for buttons and labels detection and recognition. With pairwise button detection and recognition, the framework gives a 6% higher accuracy than separate button detection and recognition. For the elevator button label recognition of separate button detection and recognition model, the accuracy score is higher than the elevator button identification step since there are multiple labels for a single button in the testing dataset. The lower performance of label detection and recognition affects the overall accuracy of the proposed method.

Fig. 9(f)-(g) show the groundtruths 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 the system while the blind user can locate the button by touching the corresponding label panel. Examples show that the control panels outside elevators for requesting up and down can be detected and recognized (the first two rows of Fig. 9). The square buttons with label symbols, the round buttons with connected label panels, the round buttons with separate label panels also can be detected by the proposed model (the last three rows of Fig. 9).

Fig. 10(a)-(c) and Fig. 11(a)-(c) show the examples of outside and indoor elevator panels for ground-truths and the final results accordingly considering button and panel as a whole. The experimental results show that most buttons and panels are detected by combining the button and panel. Notice that for the last row of Fig. 11, two bounding boxes are detected as the same button and panel and recognize both the symbol “star” button pair and “ground” button pair. This is acceptable for the cases due to the fact that the results refer to the same button with similar text representation.

Fig. 10. Elevator Button Label Recognition for outside elevator panels. (a) Input images. (b) Ground-truths of button detection. (c) Results of button recognition.

Fig. 11. Elevator Button Label Recognition. (a) Input images. (b) Ground-truths of button detection. (c) Results of button recognition.

5. CONCLUSION

To assist blind navigation, two cascade frameworks are proposed by combining with the object semantic segmentation with text recognition to detect and recognize elevator buttons and labels. The results of separate button detection and recognition are very promising and the accuracy is 73%. The pair-wised button and label detection and recognition method further proposed and boosts the accuracy up to 79.2%. The proposed method can 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

Table 2. Results for button and label detection and recognition.

Model	Step	Accuracy
	Pre-selection	92.32%
Separate Button Detection and Recognition	Elevator Button Detection	74.2%
	Elevator Button Label Recognition	71.9%
	Elevator Button Identification	73%
Pair-wise Button Detection and Recognition	-	79.2%

the queried elevator button, and evaluating the developed system and interface by blind users.

6. ACKNOWLEDGEMENTS

This work was supported in part by the National Science Foundation under award number IIS-1400802.

References

[1] Wendy P Fernandcz, Yang Xian, and Yingli Tian. Image-based barcode detection and recognition to assist visually impaired persons. In *2017 IEEE 7th Annual International Conference on CYBER Technology in Automation, Control, and Intelligent Systems (CYBER)*, pages 1241–1245. IEEE, 2017.

[2] YingLi Tian, Xiaodong Yang, Chucai Yi, and Aries Arditi. Toward a computer vision-based wayfinding aid for blind persons to access unfamiliar indoor environments. *Machine vision and applications*, 24(3):521–535, 2013.

[3] Xuejian Rong, Bing Li, J Pablo Munoz, Jizhong Xiao, Aries Arditi, and Yingli Tian. Guided text spotting for assistive blind navigation in unfamiliar indoor environments. In *International Symposium on Visual Computing*, pages 11–22. Springer, 2016.

[4] Romedi Passini and Guyltne Proulx. Wayfinding without vision: An experiment with congenitally totally blind people. *Environment and Behavior*, 20(2):227–252, 1988.

[5] Karen Duarte, Jose Cecilio, Jorge Sá Silva, and Pedro Furtado. Information and assisted navigation system for blind people. In *Proceedings of the 8th International Conference on Sensing Technology*, pages 470–473, 2014.

[6] António Pereira, Nelson Nunes, Daniel Vieira, Nuno Costa, Hugo Fernandes, and João Barroso. Blind guide: an ultrasound sensor-based body area network for guiding blind people. *Procedia Computer Science*, 67:403–408, 2015.

[7] Rai Munoz, Xuejian Rong, and Yingli Tian. Depth-aware indoor staircase detection and recognition for the visually impaired. In *2016 IEEE international conference on multimedia & expo workshops (ICMEW)*, pages 1–6. IEEE, 2016.

[8] J Pablo Muñoz, Bing Li, Xuejian Rong, Jizhong Xiao, Yingli Tian, and Aries Arditi. An assistive indoor navigation system for the visually impaired in multi-floor environments. In *2017 IEEE 7th Annual International Conference on CYBER Technology in Automation, Control, and Intelligent Systems (CYBER)*, pages 7–12. IEEE, 2017.

[9] Cang Ye and Xiangfei Qian. 3-d object recognition of a robotic navigation aid for the visually impaired. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 26(2):441–450, 2017.

[10] Cang Ye, Soonhac Hong, Xiangfei Qian, and Wei Wu. Co-robotic cane: A new robotic navigation aid for the visually impaired. *IEEE Systems, Man, and Cybernetics Magazine*, 2(2):33–42, 2016.

[11] Long Tian, Chucai Yi, and Yingli Tian. Detecting good quality frames in videos from mobile camera for blind navigation. *Journal of Computer Vision and Image Processing*, 5(10), 2015.

[12] Shuihua Wang, Chucai Yi, and Yingli Tian. Signage detection and recognition for blind persons to access unfamiliar environments. *Journal of Computer Vision and Image Processing*, 2(2), 2012.

[13] Robert Wall Emerson. Outdoor wayfinding and navigation for people who are blind: accessing the built environment. In *International Conference on Universal Access in Human-Computer Interaction*, pages 320–334. Springer, 2017.

[14] Tobias Schwarze, Martin Lauer, Manuel Schwaab, Michailas Romanovas, Sandra Böhm, and Thomas Jürgensohn. A camera-based mobility aid for visually impaired people. *KI-Künstliche Intelligenz*, 30(1):29–36, 2016.

[15] Kristof Van Beeck, Toon Goedemé, and Tinne Tuytelaars. Towards an automatic blind spot camera: robust real-time pedestrian tracking from a moving camera. In *Proceedings of the twelfth IAPR conference on machine vision applications*, pages 528–531. MVA Organization; Japan, 2011.

[16] Hugo Fernandes, Vitor Filipe, Paulo Costa, and João Barroso. Location based services for the blind supported by rfid technology. *Procedia Computer Science*, 27:2–8, 2014.

[17] Saleh Alghamdi, Ron Van Schyndel, and Ahmed Alahmadi. Indoor navigational aid using active rfid and qr-code for sighted and blind people. In *2013 IEEE Eighth International Conference on Intelligent Sensors, Sensor Networks and Information Processing*, pages 18–22. IEEE, 2013.

[18] Lorenzo Picinali, Amandine Afonso, Michel Denis, and Brian FG Katz. Exploration of architectural spaces by blind people using auditory virtual reality for the construction of spatial knowledge. *International Journal of Human-Computer Studies*, 72(4):393–407, 2014.

[19] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

[20] Juan Manuel Sáez, Francisco Escolano, and Miguel Angel Lozano. Aerial obstacle detection with 3-d mobile devices. *IEEE journal of biomedical and health informatics*, 19(1):74–80, 2014.

[21] Dong Zhang, Dah-Jye Lee, and Brandon Taylor. Seeing eye phone: a smart phone-based indoor localization and guidance system for the visually impaired. *Machine vision and applications*, 25(3):811–822, 2014.

[22] Catherine Feng, Shiri Azenkot, and Maya Cakmak. Designing a robot guide for blind people in indoor environments.

In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts*, pages 107–108. ACM, 2015.

[23] Department of Justice. Ada standards for accessible design. *Title III regulation*, 28, 2010.

[24] Jingya Liu and Yingli Tian. Recognizing elevator buttons and labels for blind navigation. In *2017 IEEE 7th Annual International Conference on CYBER Technology in Automation, Control, and Intelligent Systems (CYBER)*, pages 1236–1240. IEEE, 2017.

[25] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[26] Jifeng Dai, Kaiming He, and Jian Sun. Instance-aware semantic segmentation via multi-task network cascades. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3150–3158, 2016.

[27] Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. Segmentation from natural language expressions. In *European Conference on Computer Vision*, pages 108–124. Springer, 2016.

[28] Minghui Liao, Baoguang Shi, Xiang Bai, Xinggang Wang, and Wenyu Liu. Textboxes: A fast text detector with a single deep neural network. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[29] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[30] Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip HS Torr. Conditional random fields as recurrent neural networks. In *Proceedings of the IEEE international conference on computer vision*, pages 1529–1537, 2015.

[31] Pierre Baqué, Timur Bagautdinov, François Fleuret, and Pascal Fua. Principled parallel mean-field inference for discrete random fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5848–5857, 2016.

[32] Baoguang Shi, Xiang Bai, and Cong 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*, 39(11):2298–2304, 2016.