Computers, Materials & Continua
DOI: 10.32604/cmc.2023.033327
Article

Tech Science Press

An Elevator Button Recognition Method Combining YOLOv5 and OCR

Xinliang Tang[1], Caixing Wang[1], Jingfang Su[1,*] and Cecilia Taylor[2]

[1]School of Information Science and Engineering, Hebei University of Science and Technology, Shijiazhuang, 050000, China
[2]Nankai-Birmingham Institute of Data Science Intelligence, Birmingham, B100AB, Britain
*Corresponding Author: Jingfang Su. Email: sujingfang1980@hebust.edu.cn
Received: 14 June 2022; Accepted: 15 November 2022

Abstract: Fast recognition of elevator buttons is a key step for service robots to ride elevators automatically. Although there are some studies in this field, none of them can achieve real-time application due to problems such as recognition speed and algorithm complexity. Elevator button recognition is a comprehensive problem. Not only does it need to detect the position of multiple buttons at the same time, but also needs to accurately identify the characters on each button. The latest version 5 of you only look once algorithm (YOLOv5) has the fastest reasoning speed and can be used for detecting multiple objects in real-time. The advantages of YOLOv5 make it an ideal choice for detecting the position of multiple buttons in an elevator, but it's not good at specific word recognition. Optical character recognition (OCR) is a well-known technique for character recognition. This paper innovatively improved the YOLOv5 network, integrated OCR technology, and applied them to the elevator button recognition process. First, we changed the detection scale in the YOLOv5 network and only maintained the detection scales of $40 * 40$ and $80 * 80$, thus improving the overall object detection speed. Then, we put a modified OCR branch after the YOLOv5 network to identify the numbers on the buttons. Finally, we verified this method on different datasets and compared it with other typical methods. The results show that the average recall and precision of this method are 81.2% and 92.4%. Compared with others, the accuracy of this method has reached a very high level, but the recognition speed has reached 0.056 s, which is far higher than other methods.

Keywords: Button recognition; deep learning; multi-object detection

1 Introduction

Elevator button recognition is one of the most critical functions in the process of robot delivery to the home. In the process of robot delivering goods, the movement between floors is mainly completed by taking the elevator, while the navigation research of the robot is mainly based on plane path planning, and little attention is paid to the robot autonomously taking the elevators. It is of great significance for robots to have the ability to take the elevator independently to complete the whole delivery process end to end. Therefore, how to make the robot take the elevator autonomously becomes an important problem [1].

The process of taking the elevator autonomously for the robot involves key technologies such as object detection and mechanical control. To give robots the ability to take the elevator, domestic and foreign researchers have carried out a large amount of research. The main ideas are divided into two kinds. The first one is to use the chip scheduling robot. The operation of this kind of robot in the elevator mainly depends on the chip embedded in the robot and the elevator's central control system. This method requires customization of the elevator scene and the robot, which makes the robot not universal in the application scene. The second one, which has been widely studied, is that the robot relies on its computer vision capabilities, but it requires the robot to run a certain object detection algorithm locally. In the following part, we focus on how to use the object detection algorithm in the field of computer vision to complete elevator button recognition.

More object detection algorithms based on deep learning [2] are applied in hardware because of the emergence of deep learning, the continuous evolution of hardware devices, and the continuous strengthening of computing power. Compared with traditional methods, deep learning mechanisms can significantly improve object detection performance and have strong anti-interference ability. Thus, the object detection algorithm based on deep learning has gradually been widely used in unmanned aerial vehicles (UAV), autonomous driving, service robots, and other fields. With the development of deep learning, great breakthroughs have been made in the field of object detection, especially in the development of extraction algorithms of convolutional neural networks (CNN) and regions of interest (ROI). Many researchers applied it to button recognition for robots taking the elevator.

The object detection algorithm based on deep learning mainly adopts the network structure of the convolutional neural network, which is divided into a two-step algorithm and a single-step algorithm according to the completion of the object detection task. Region-based convolutional neural networks (RCNN), fast RCNN, and faster RCNN [3–5] belong to two-step object detection algorithms, which complete the object detection in two stages. Firstly, candidate regions are selected, and then these candidate regions are sent to the network structure to extract features, which can predict the location and category of the detected objects. The two-step algorithm takes a lot of time to generate region proposals, so the detection speed is relatively slow and difficult to apply to the field of real-time detection.

The typical representative of the single-step object detection algorithm is you only look once (YOLO) algorithm proposed by Redmon et al. [6] in 2016 to improve the detection speed. This algorithm does not need to generate candidate regions in advance but can identify the categories and positions of multiple items in one image at one time, achieving end-to-end image recognition, and greatly improving the computing speed. Therefore, this kind of algorithm has been widely used due to its simple deployment and fast detection speed. The current problem is that the detection accuracy of small objects is not ideal. At present, the YOLO algorithm has experienced development from version 1 (YOLOv1) to version 5 (YOLOv5), and the algorithm performance is gradually improved. Especially in 2020, Ultralights launched the YOLOv5 algorithm [7], which uses mosaic data enhancement and multi-scale detection function to effectively make up for the defect of YOLO's insensitivity to small objects, and has the advantages of small size, fast speed, and high accuracy.

YOLOv5 has been well used in traffic scenarios, but few have tried to use it in elevator button recognition. Due to this small size advantage, the YOLOv5 algorithm is very suitable for robots to recognize multiple elevator buttons in an elevator environment. This paper mainly studies the recognition technology of elevator buttons. Based on the comprehensive consideration of the weight file size, recognition precision, and detection speed, the YOLOv5 algorithm is combined with the optical character recognition (OCR) algorithm, which can improve the recognition performance of the elevator buttons [8].

The paper is organized as follows: Section 2 introduces the research status of elevator button recognition. Section 3 introduces the details of the algorithm and the network structure used in this paper. Section 4 introduces the preparation process of elevator button datasets and the experimental process of elevator button recognition, and compares the experimental results with other methods. The conclusion can be found at the end of this paper.

2 Related Works

Traditional button recognition methods often contain a two-stage process, first detecting the buttons and then recognizing the text on the buttons by OCR algorithm or classification networks. For example, Klingbeil et al. [9] used the two-dimensional sliding window detector for the preliminary detection of the elevator buttons and applied the expectation maximization (EM) algorithm and grid model to adjust the button position to correct the excessive or missed detection. After that, OCR technology was used to binarize and recognize the label of the elevator buttons. Although the method can accurately recognize the elevator buttons, the grid model needs to have appropriate rows and columns to correctly match each panel, so the algorithm allows robots to learn up to five grid models in the process of machine learning, which makes the whole recognition process time-consuming and inefficient. Kim et al. [10] first detected the button position and label region of interest through feature extraction, object segmentation, and object registration, and then used a template-matching algorithm for label recognition and conducted a refinement. The method can improve the performance of object segmentation in images and solve the ambiguity problem introduced by simple or repeated patterns on the buttons, but it has specific requirements for the shape of the elevator buttons, that is, only the buttons with convex quadrangle boundary are applicable. Islam et al. [11] used the histogram of oriented gradient (HOG) algorithm to extract features of the elevator button images and the bag of words (BOW) method to extract features of the elevator button images, and then used the image classification of artificial neural network (ANN) to perform the object identification [12]. This method is accurate but time-consuming. Due to the limited computing resources of service robots and the competition of some other programs for computing resources, the two-stage methods based on these deep learning networks and OCR cannot be directly applied to real-time object detection of robots.

Later, some researchers directly applied the end-to-end multi-object detection framework to the detection of elevator buttons and text recognition. For example, Dong et al. [13] proposed a recognition model based on a convolutional neural network (CNN), which combined an ROI extraction algorithm with button arrangement rules to achieve simultaneous location and recognition of elevator buttons. Liu et al. [14] used a single detector (SSD) to extract the text description and spatial location of the button label, and then used the full convolutional network (FCN) to estimate the semantic pixel-level semantic segmentation mask for elevator button recognition. Yang et al. [15] use multi-object detection neural network to detect and classify buttons simultaneously. A framework of YOLO is used to perform end-to-end recognition in this work. Although only a limited number of button categories can be recognized, the robustness of this method is significantly improved. All the above-mentioned end-to-end methods have certain operational stability in elevator button recognition, but the detection accuracy and processing time still need to be improved, so they cannot be directly used for real-time detection.

In recent research, Wang et al. [16] introduced the idea of an Inception module and used convolution kernels of different sizes to operate in parallel, increasing the depth and width of networks at all levels, and improving the ability of network feature extraction. He also combines batch normalization (BN) algorithm to improve the model training speed and network classification

ability. The accuracy of the improved algorithm is 2% higher than that of the original algorithm. Liu et al. [17] successfully used the OCR model implemented through the RCNN network to recognize the characters of elevator buttons to achieve the recognition process of elevator buttons. Meanwhile, they released the first large-scale public elevator panel dataset, which aims to provide a benchmark for the segmentation and recognition methods of elevator buttons in the future.

Some current deep learning algorithms have achieved some visible results in object recognition applications, which fully show that the application of deep learning algorithms in elevator button recognition is feasible. Nevertheless, the process of quickly and accurately recognizing all elevator buttons under complex backgrounds and different light reflection conditions still needs to be improved [18].

OCR [19] is a technology that converts the text in images into a text format in an optical way through a series of image processing and recognition algorithm. Binarization, noise reduction, and other techniques in OCR can effectively solve the shortcomings of elevator interiors, and are suitable for key text recognition [20]. Therefore, this paper combines the YOLOv5 algorithm with OCR technology to improve the detection speed and reduce the impact of the complex environment inside the elevator as much as possible thus improving the overall detection efficiency and accuracy. This method is called the YOLOv5 + OCR model in this paper.

3 Research Method

3.1 Object Detection Network

YOLO is a fast and compact open-source object detection model, with faster recognition speed, stronger performance, and good stability at the same scale compared with other networks. It is an end-to-end neural network that can predict object categories and bounding boxes.

YOLOv1 classifies the generation of region proposal and feature extraction into one step, divides the image into S × S grids, and extracts features on the grids, each grid corresponds to two bounding boxes. The speed is improved, but some objects will be lost. YOLOv2's backbone uses darknet-19, removing the full connection layer and using five anchor boxes on each grid to detect some objects that have not yet been learned. It can identify more objects and improve speed and accuracy but it is still not accurate in small object detection [21]. YOLOv3 uses the new backbone network, in which the head partially uses multi-scale features to detect objects of different sizes simultaneously. YOLOv4 incorporates several improvements over YOLOv3, using mosaic data augmentation, cross-small batch normalization, and self-adversarial training (SAT) in the input. Its backbone network adopts the cross-stage partial network (CSP) and selects the genetic algorithm to prevent over-fitting. The head part still uses the multiscale feature detection in YOLOv3 [22].

YOLOv5 also performs mosaic data augment operations for image input and further uses adaptive scaling operations for inference. This method can adaptively fill according to different input image sizes, improving the inference speed by 37%. Secondly, the FOCUS structure located at the front of the network is designed. The main content of this structure is to slice the input data, which can effectively improve the quality of image feature extraction. In the output, YOLOv5 can better detect some overlapping objects in the detected image without increasing computational resources. According to the official data, the fastest reasoning time per image in the current version of YOLOv5 is 0.007 s, or 140 frames per second (FPS), and the weight file size is only 1/9 of YOLOv4 [23].

YOLOv5 comes in four weight models, namely YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x. The comparison of the four models is shown in Fig. 1. YOLOv5s is the fastest-weight model with the

smallest feature mapping width in this series. Although the accuracy is not the highest, it can also meet our recognition requirements after subsequent improvement. Therefore, we choose the elevator button recognition experiment based on YOLOv5s.

The network structure of YOLOv5 is mainly composed of four parts. Their functions are listed below. [24]

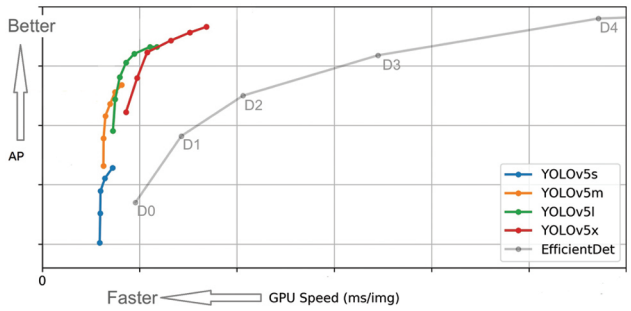


Figure 1: Comparison of four weight models

Input: Complete adaptive anchor frame calculation and adaptive image scaling.

Backbone: Convolutional neural network for feature extraction of images.

Neck: Combine and mix the image features, and pass the results to the prediction layer.

Head: Produce the final prediction output of the location and category.

3.2 The Improved Recognition Network

3.2.1 The Improved YOLOv5 Network

The original YOLOv5s has three detection scales: 20×20, 40×40, and 80×80. The detection scale of 20×20 is mainly used to detect large objects, the detection scale of 40×40 is mainly used to detect medium-sized objects and the detection scale of 80×80 is used to detect small objects. [25]

Since the button detection in the elevator scene is a small detection object, the large objects detection scale of 20×20 is removed from the three detection scales in our method. If only the detection scale of 80×80 is retained, some datasets may fail to extract effective features due to unequal shooting distances and uneven pixel distribution, thus affecting the training results. Therefore, we chose to retain a 40×40 medium size and 80×80 small size detection scale. The network can better meet the requirements of the detected objects and further improve the network speed by doing this. The improved network structure is shown in Fig. 2.

The operation process of the YOLOv5 network is as follows. The input part is used as the input of the dataset. Firstly, the image features of the dataset are extracted by the backbone part, and the image is segmented into 4 images with the size of $128 \times 128 \times 3$ according to focal length. After processing, the feature map with the size of $64 \times 64 \times 12$ can be obtained. The focus saves the result of the multiple-slices feature maps, and then sends them to the module, where the convolution operation is used to extract features from the feature map. After each convolution operation, a concentrated-comprehensive convolution block (C3) is used to reduce the computation and memory. Each C3 in the backbone is followed by a spatial attention module (SAM) to extract key features and suppress irrelevant information.

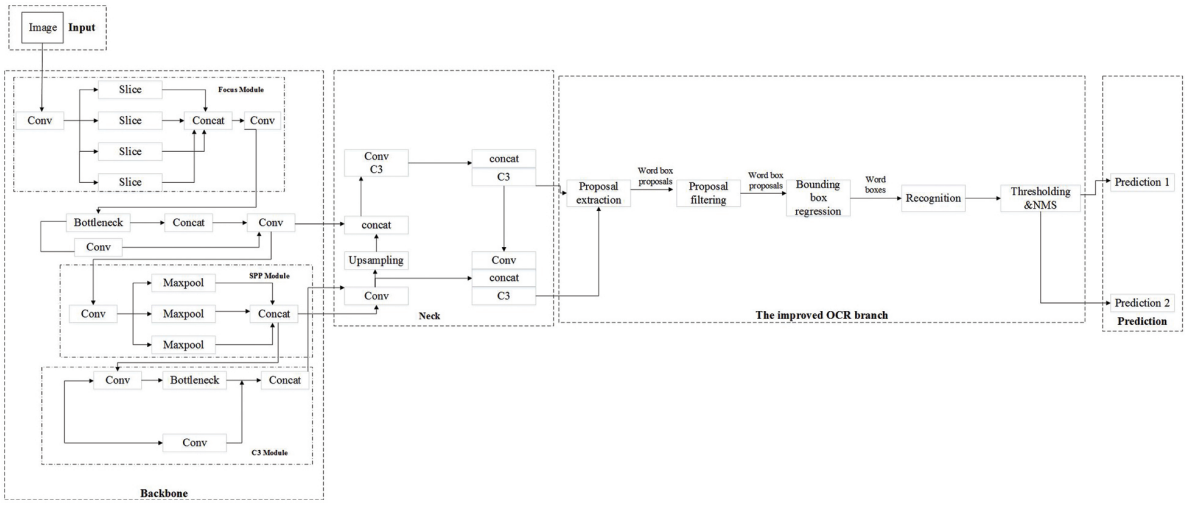


Figure 2: The improved YOLOV5 network structure

The neck part adopts the feature pyramid network (FPN) structure to transfer high-level features from top to bottom, enhances semantic features, and improves the feature extraction ability of the network. As shown in Eq. (1), the prediction part uses GIOULOSS to perform the loss function of the bounding box and filters the object box by non-maximum suppression.

$$\text{GIOULOSS} = 1 - IOU + \frac{\rho^2(b, b^{gt})}{c^2} \tag{1}$$

where b is the center point of the dimension box, b^{gt} is the regression center, $\rho()$ is the Euclidean distance between the center point of the dimension box and the center point of the regression box, c is the diagonal length of the smallest circumscribed rectangle between the dimension box and the regression box, IOU is the ratio of the intersection set and union set of dimension box and regression box.

3.2.2 The Combination of OCR Technology

The elevator button position detection can be well solved by using the YOLOv5s object detection algorithm. However, YOLOv5's capabilities are still very limited when it comes to specific word recognition, especially when most elevator panels have mirror reflections. Simply describing the recognition task as a multi-object detection problem will limit the improvement of recognition performance. Therefore, the YOLOv5 is mainly used for position detection of elevator buttons in our work, and the text recognition on the buttons is carried out by the OCR algorithm after the accurate position of the buttons is given.

The main techniques used in our approach include the following sections: [26]

Binarization: Most of the images taken by the camera are colorful, and contain a huge amount of information. To make the computer recognize characters faster and better, colorful images are processed into the images containing only foreground information and background information, the foreground information is simply defined as black, and also, the background information is white.

Denoising: For different images, the definition of noise is different. Denoising is carried out according to the characteristics of noise.

Tilt correction: Images taken by the camera are inevitably tilted, so correction of tilt is required.

Character segmentation: Because of the limitations of illumination, specular reflection, and other conditions during photographing, the captured images may have problems of character adhesion or incomplete character, which greatly limit the performance of the recognition model. Thus, a character segmentation operation is required for the captured images.

Character recognition: Character recognition process is realized through dataset training, character feature extraction and template matching.

This experiment mainly uses the binarization, denoising, character segmentation, and character recognition processes of OCR technology, and combines the characteristics of the YOLOv5 neural network that can be used to improve the object recognition speed and character recognition accuracy, to achieve a complete elevator button recognition process. The main process of the YOLOv5 + OCR model is shown in Fig. 3.

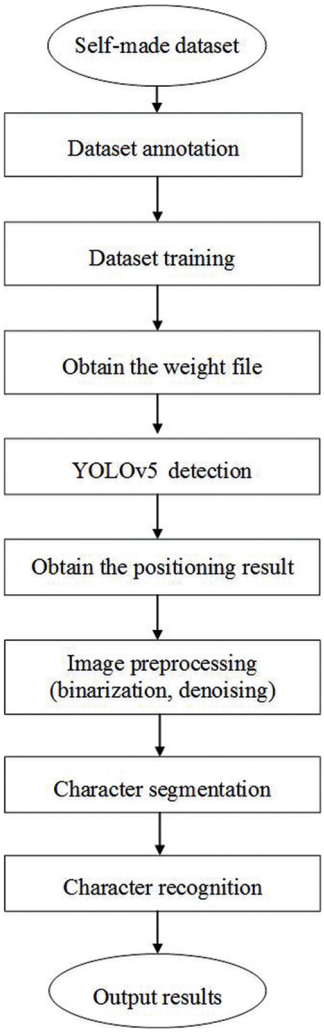


Figure 3: Flow chart of elevator button recognition model

4 Experimental Process and Results

To verify the effectiveness of our proposed approach, we collected a large dataset containing exterior and interior images of elevator button panels. The dataset contains a total of 102 different elevator panels, including 478 button labels and 46 different categories. These data come from network searches and collections in daily life. Once enough datasets had been collected, the next step was to add labels on the data images. After converting the label files to text files, we input them into the YOLOv5 network for training.

4.1 Button Detection

As shown in Fig. 2, the YOLOv5 network structure includes a primary prediction box. Its core idea is to take the whole image as the input of the network and directly predict the position and category of the object box in the output layer. An object box corresponds to four location information and one confidence information. The confidence equation is shown in Eq. (2).

$$\mathrm{P_r}\,(\text{Classi Object}) * \mathrm{P_r}\,(\text{Object}) * \mathrm{IOU}\frac{\text{truth}}{\text{pred}} = P_r\,(Classi) * IOU\frac{truth}{pred} \tag{2}$$

The first item on the left of the equation is the category probability of each grid prediction, and the second and third items are the confidence of each object box prediction. The result of their multiplication contains not only the probability that the predicted object box belongs to a certain category but also the probability of the accuracy of the object box. After obtaining the category confidence score of each object frame, the low-score object frames are filtered out by comparing them with the threshold. The retained object frames are the final detection results after non-maximum suppression (NMS) processing.

Firstly, the elevator button image captured by the camera on the robot is sent to our trained network model for button positioning and bounding box recognition. Secondly, these located object areas are passed to the customized OCR engine one by one. After reading these areas, the OCR engine stores them and runs the template matching algorithm to get the character recognition results, and then outputs the floor number on the elevator buttons in the form of a label on the original image to complete the recognition task of elevator buttons. Some experimental comparison results are as follows in Fig. 4. The OCR network is a character-level recognition based on the character dictionary. Due to the serious data imbalance problem in the elevator button dataset, the multi-object detection model is easy to over-fit classes of training samples. The OCR network can decompose a large number of floor numbers into individual characters and then identify them separately, thus, the number of categories that need to be identified is greatly reduced and the problem of data imbalance is alleviated.

4.2 Loss Function

The process of training the dataset involves label classification and sample regression training, and its loss functions are respectively expressed as classification loss L_c and regression loss L_r, where Lc is a typical binary cross entropy (BCE) loss, as shown in Eq. (3).

$$\mathrm{L_c} = \frac{1}{N_c}\sum_{i=1}^{N_c} BCE\left(b_i, b_i^*\right), b_i, b_i^* \in \{0, 1\} \tag{3}$$

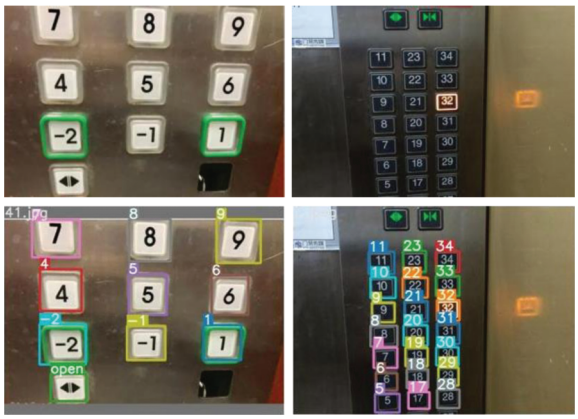


Figure 4: Experimental comparison results

Lr is the Huber loss of robust regression, as shown in Eqs. (4) and (5). N_r and N_c represent the number of button candidate boxes.

$$\mathrm{L_r} = \frac{1}{N_r}\sum_{i=1}^{Nr} 1\left[b_i \in \{1\}\right] Huber\left(p_i, p_i^*\right) \tag{4}$$

$$\mathrm{Huber}\left(\mathrm{p_i}, \mathrm{p_i}^*\right) = \begin{cases} 0.5||p - p^*||_2^2, ||p - p^*||1 < 1 \\ ||p - p^*||1 - 0.5, otherwise \end{cases} \tag{5}$$

The training can be stopped when the average loss is no longer reduced after a certain number of iterations. Then the weight file with the highest score will be selected as the network weight file.

In the process of object detection, the optical character recognition loss L_o is shown in Eq. (6). L_o is the loss of case level classification cross entropy (CCE). The loss function of the whole process can be expressed as Eq. (7).

$$L_o = \frac{1}{N_o}\sum_{i=1}^{No} 1\left[b_i \hat{\mathrm{I}}\{1\}\right]\sum_{t=1}^{T} CCE\left(o_i^t, c_i^t\right) \tag{6}$$

$$\mathrm{L} = \lambda cLc + \lambda_r L_r + \lambda_o L_o \tag{7}$$

By defining and optimizing the loss function, the network model can improve recognition accuracy to a certain extent. Compared with other models, our model has less loss and a better effect. The comparison of loss functions of different models is shown in Fig. 5.

4.3 Mean Average Precision (mAP)

To further explore the feasibility of YOLOv5 in the application of elevator button recognition, this section compares the mean average precision (mAP) of different object detection algorithms in the application of elevator button recognition in recent years, as shown in Fig. 6.

As can be seen from Fig. 6, although the mAP of YOLOv5 is not the highest, it is not much different from the RCNN algorithm with the highest value. This disadvantage can be remedied by combining it with OCR. In addition, the ultra-high recognition speed of the YOLOv5 algorithm completely opens a new world in the field of elevator button recognition.

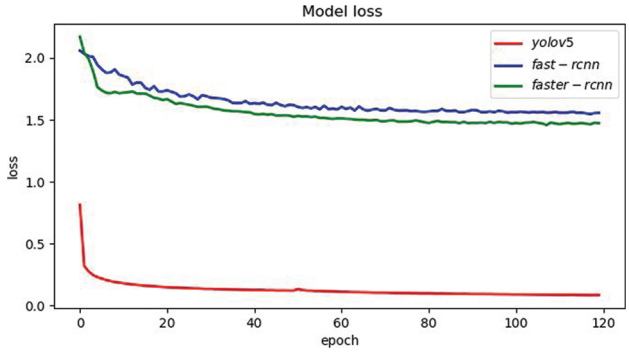


Figure 5: Comparison of loss functions of different models

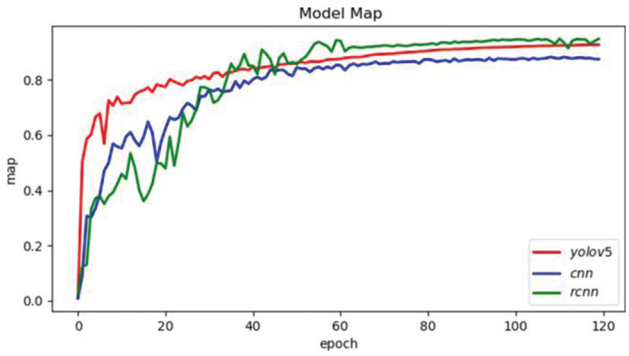


Figure 6: mAP value comparison of different models

4.4 Results and Analysis

In this section, we first analyzed the overall performance of the YOLOv5 + OCR model in the process of elevator button recognition based on the experimental results. Then we made a comprehensive comparison of the proposed method with other existing detection methods and discussed the results.

Firstly, the following performance indicators are used to evaluate the performance of the object detection algorithm: precision (P), recall (R), and mAP. P is a measure of accuracy and R is a measure of coverage, as shown in Eqs. (8) and (9).

$$Precision = TP/(TP + FP) \times 100\% \tag{8}$$

$$\text{Recall} = TP/(TP + FN) \times 100\% \tag{9}$$

where TP represents the number of predicted positive cases that are positive, FP represents the number of predicted positive cases that are negative and FN represents the number of predicted negative cases that are positive. The mAP is used to measure recognition accuracy, which is obtained by averaging the average accuracy values of all categories.

After the training of YOLOv5 network label classification and location detection, the maximum P, R, mAP_0.5 and mAP_0.5:0.95 of the dataset respectively reach 0.948, 0.989, 0.996 and 0.958, as shown in Fig. 7 and the dataset analysis diagrams are shown in Fig. 8.

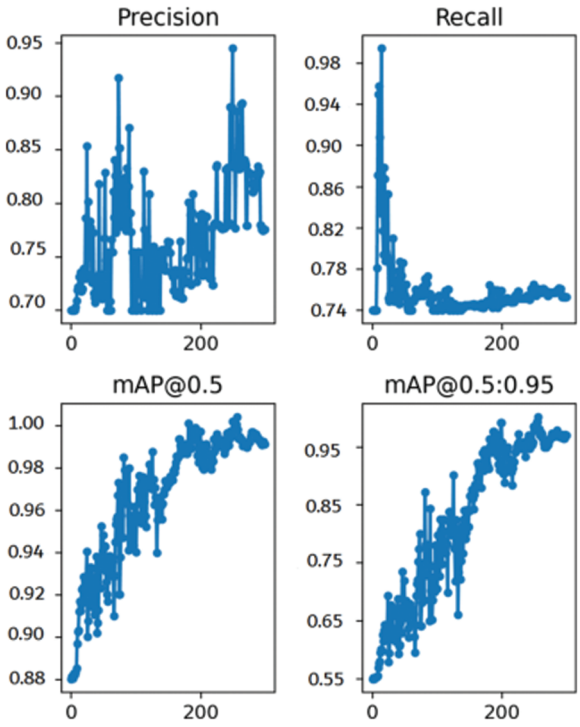
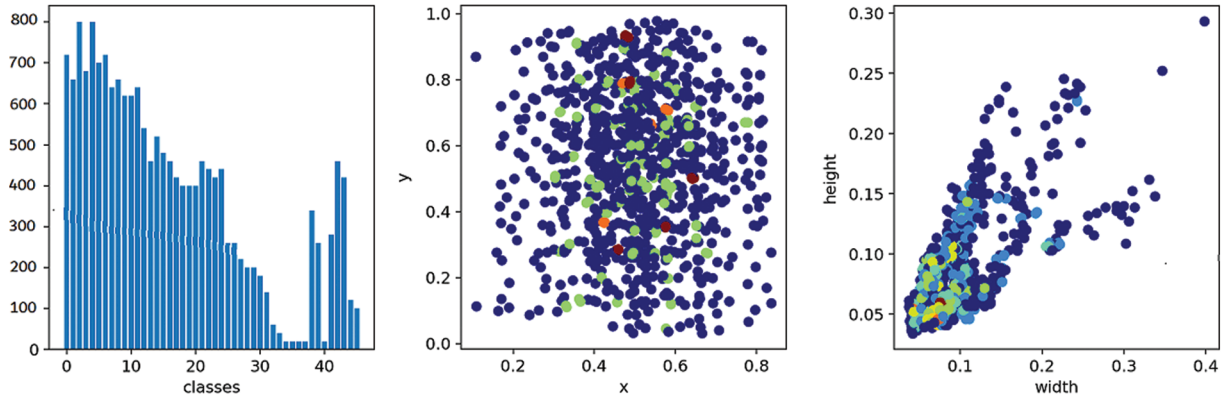


Figure 7: Model performance evaluation index



(a) Dataset category classification. (b) Dataset objects center position distribution. (c) Dataset object size segment.

Figure 8: Dataset analysis

To verify the accuracy of the YOLOv5 + OCR model in elevator button recognition, we compared the recognition results of the YOLOv5 + OCR model on self-made datasets trained by ourselves and other two different datasets. One of the datasets comes from collection images on the website; the other comes from a part of the first publicly available large-scale elevator key panel dataset published by Zhu et al. [1]. The test results of the YOLOv5 + OCR model on different datasets are shown in Table 1.

Table 1: Test results of different elevator button datasets under the YOLOv5 + OCR model. AR represents average recall, AP represents average precision and the time represents the processing time of each test image

Test set type	Size test set		Test result		
	Number of buttons	Label category	AR	AP	Time/s
Network dataset	485	37	0.719	0.931	0.052
Reference dataset	492	50	0.778	0.944	0.046
Self-made dataset	478	46	0.706	0.926	0.056

Through the above experimental results, it is obvious that the difference in AP between the trained self-made dataset and the other two datasets is little, which shows that the YOLOv5 + OCR model is not only applicable to our dataset but also widely applicable to other datasets of elevator buttons in a variety of scenarios.

To further study the performance of our proposed model and other existing elevator button recognition methods, such as the traditional template-matching model, traditional CNN + OCR model added expectation-maximum algorithm (EM) and hidden Markov model (HMM), end-to-end CNN model and OCR-RCNN model. We analyzed another series of experiments about these methods on the datasets we constructed above. The experimental results are shown in Table 2 below. The test results are affected by the input image resolution, image acquisition speed, model size, and so on.

Table 2: Performance comparison of different elevator button recognition methods

Experimental method	Test set type	Size test set		Test result		
		Number of buttons	Label category	AR	AP	Processing time/s
Traditional template-matching model	Self-made dataset	478	46	—	—	0.380
Traditional CNN + OCR(added EM and HMM)				0.855	0.880	3.000
End-to-end CNN				0.800	0.876	0.417
OCR-RCNN				0.815	0.945	0.142
YOLOv5 + OCR model				0.812	0.924	0.056

From the experimental results of different elevator button recognition methods in Table 2, we could conclude that the AP of the OCR-RCNN model and YOLOv5 + OCR model are higher than others and the accuracy of both models remains at a high level, with little difference between them. However, the recognition speed of the YOLOv5 + OCR model is much higher than that of the OCR-RCNN model. Besides, it is not difficult to see that the YOLOv5 + OCR model adopted in this paper

has shown great advantages in the experiment process of elevator button recognition. It not only has a strong recognition speed, but also effectively solves the problem of data imbalance, and it has a great application value in reality.

5 Conclusions

In this paper, an elevator button recognition method based on the YOLOv5 and OCR is proposed. In the object detection stage, we adopted the YOLOv5 algorithm to complete the deep learning task, which greatly improved the overall detection speed. To verify the effectiveness of the YOLOv5 + OCR model, we collected a large elevator button dataset and conducted template matching training. Compared with many other object recognition models, the YOLOv5 + OCR model has obvious advantages in running speed, better performance of detecting small objects, and clearer results. The results of the experiment inferred that the YOLOv5 + OCR model can be well applied to the field of service robots that take elevators autonomously and can recognize elevator buttons quickly and accurately. It is of great practical significance to improve the speed and accuracy of elevator button recognition for robots riding elevators autonomously.

Acknowledgement: We acknowledge funding from the Research and Implementation of An Intelligent Driving Assistance System Based on Augmented Reality in Hebei Science and Technology Support Plan (Grant Number 17210803D), Science and Technology Research Project of Higher Education in Hebei Province (Grant Number ZD2020318), Middle School Students Science and Technology Innovation Ability Cultivation Special Project (Grant No.22E50075D) and project (Grant No. 1181480).

Funding Statement: This research was funded by the Research and Implementation of An Intelligent Driving Assistance System Based on Augmented Reality in Hebei Science and Technology Support Plan (Grant Number 17210803D), Science and Technology Research Project of Higher Education in Hebei Province (Grant Number ZD2020318), Middle School Students Science and Technology Innovation Ability Cultivation Special Project (Grant No.22E50075D) and project (Grant No. 1181480).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] D. Zhu, Y. Fang, Z. Min, D. Ho and Q. H. Meng, "OCR-RCNN: An accurate and efficient framework for elevator button recognition," *IEEE Transactions on Industrial Electronics*, vol. 69, no. 1, pp. 582–591, 2021.

[2] W. H. Wang and J. Y. Tu, "Research on license plate recognition algorithms based on deep learning in complex environment," *IEEE Access*, vol. 8, pp. 91661–91675, 2020.

[3] X. R. Zhang, J. Zhou, W. Sun and S. K. Jha, "A lightweight CNN based on transfer learning for COVID-19 diagnosis," *Computers, Materials & Continua*, vol. 72, no. 1, pp. 1123–1137, 2022.

[4] S. Zhao, Z. L. Liu, A. L. Zheng and Y. Gao, "Real-time classification and detection of garbage based on SSD improved with mobilenetv2 and IFPN," *Computer Application*, vol. 42, no. 1, pp. 1–10, 2022.

[5] H. Jiang and E. Learned-Miller, "Face detection with the faster R-CNN," in *2017 12th IEEE Int. Conf. on Automatic Face & Gesture Recognition (FG 2017)*, Washington, DC, USA, pp. 650–657, 2017.

[6] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You only look once: Unified, real-time object detection," in *2016 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, USA, pp. 779–788, 2016.

[7] G. Jocher, A. Stoken, J. Borovec, N. Code, A. Chaurasia *et al.,* "Ultralytics/yolov5: V5.0-YOLOv5-p6 1280 models," AWS, Supervise.ly and YouTube integrations, 2021. https://doi.org/10.5281/zenodo.4679653.

[8] W. Sun, X. Chen, X. R. Zhang, G. Z. Dai, P. S. Chang *et al.,* "A Multi-feature learning model with enhanced local attention for vehicle re-identification," *Computers, Materials & Continua*, vol. 69, no. 3, pp. 3549–3560, 2021.
[9] E. Klingbeil, B. Carpenter, O. Russakovsky and A. Y. Ng, "Autonomous operation of novel elevators for robot navigation," in *2010 IEEE Int. Conf. on Robotics and Automation*, Anchorage, Alaska, USA, pp. 751–758, 2010.
[10] H. H. Kim, D. J. Kim and K. H. Park, "Robust elevator button recognition in the presence of partial occlusion and clutter by specular reflections," *IEEE Transactions on Industrial Electronics*, vol. 59, no. 3, pp. 1597–1611, 2011.
[11] K. T. Islam, G. Mujtaba, R. G. Raj and H. F. Nweke, "Elevator button and floor number recognition through hybrid image classification approach for navigation of service robot in buildings," in *Int. Conf. on Engineering Technology and Technopreneurship*, Antalya, Turkey, pp. 1–4, 2017.
[12] A. A. Abdulla, H. Liu, N. Stoll and K. Thurow, "A robust method for elevator operation in semi-outdoor environment for mobile robot transportation system in life science laboratories," in *IEEE Jubilee Int. Conf. on Intelligent Engineering Systems*, Budapest, Hungary, pp. 45–50, 2016.
[13] Z. Dong, D. Zhu and M. Q. -H. Meng, "An autonomous elevator button recognition system based on convolutional neural networks," in *Robotics and Biomimetics (ROBIO), 2017 IEEE Int. Conf. on*, Macau, China, pp. 2533–2539, 2017.
[14] J. Liu and Y. Tian, "Recognizing elevator buttons and labels for blind navigation," in *2017 IEEE 7th Annual Int. Conf. on CYBER Technology in Automation, Control, and Intelligent Systems (CYBER)*, Hawaii, USA, IEEE, pp. 1236–1240, 2017.
[15] P. -Y. Yang, T. -H. Chang, Y. -H. Chang and B. -F. Wu, "Intelligent mobile robot controller design for hotel room service with deep learning armbased elevator manipulator," in *2018 Int. Conf. on System Science and Engineering (ICSSE)*, Taiwan, China, pp. 1–6, 2018.
[16] T. Y. Wang, W. Guo and Y. P. Wu, "Elevator passenger identification method based on multi task convolutional neural network," *Computer System Application*, vol. 30, no. 6, pp. 278–285, 2021.
[17] J. Liu, Y. Fang and D. Zhu, "A Large-scale dataset for benchmarking elevator button segmentation and character recognition," in *2021 IEEE Int. Conf. on Robotics and Automation (ICRA 2021)*, Xi'an, China, pp. 14018–14024, 2021.
[18] D. Zhu, T. Li, D. Ho, T. Zhou and Q. H. Meng, "A novel ocr-rcnn for rlevator button recognition," in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, Madrid, Spain, IEEE, pp. 3626–3631, 2018.
[19] T. Zhang, M. D. Ma and D. Y. Wang, "Research on OCR character recognition technology," *Computer Technology and Development*, vol. 30, no. 4, pp. 85–88, 2020.
[20] W. Sun, G. Z. Dai, X. R. Zhang, X. Z. He and X. Chen, "TBE-Net: A three-branch embedding network with part-aware ability and feature complementary learning for vehicle re-identification," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 9, pp. 14557–14569, 2021. https://doi.org/10.1109/TITS.2021.3130403.
[21] X. R. Zhang, H. L. Wu, W. Sun, A. G. Song and S. K. Jha, "A fast and accurate vascular tissue simulation model based on point primitive method," *Intelligent Automation & Soft Computing*, vol. 27, no. 3, pp. 873–889, 2021.
[22] Z. G. Li and N. Zhang, "A light weight YOLOv5 traffic sign recognition method," *Telecommunication Engineering*, vol. 62, no. 9, pp. 1201–1206, 2021. http://kns.cnki.net/kcms/detail/51.1267.tn.20211230.1940.014.html.
[23] W. Sun, L. Dai, X. R. Zhang, P. S. Chang and X. Z. He, "RSOD: Real-time small object detection algorithm in UAV-based traffic monitoring," *Applied Intelligence*, vol. 52, no. 6, pp. 8448–8463, 2021. https://doi.org/10.1007/s10489-021-02893-3.
[24] H. B. Fang, G. Wan, Z. H. Chen, Y. W. Huang, W. Y. Zhang *et al.,* "Offline handwriting mathematical symbol recognition based on improved YOLOv5s," *Journal of Graphics*, vol. 43, no. 3, pp. 387–395, 2021. https://kns.cnki.net/kcms/detail/10.1034.T.20211231.1512.004.html.

[25] S. Jiang, W. Yao, M. S. Wong, M. Hang, Z. H. Hong *et al.,* "Automatic elevator button localization using a combined detecting and tracking framework for multi-story navigation," *IEEE Access*, vol. 8, pp. 1118–1134, 2019.
[26] W. Sun, G. C. Zhang, X. R. Zhang, X. Zhang and N. N. Ge, "Fine-grained vehicle type classification using lightweight convolutional neural network with feature optimization and joint learning strategy," *Multimedia Tools and Applications*, vol. 80, no. 20, pp. 30803–30816, 2021.