International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 14 Issue 5 ǁ May 2026 ǁ PP. 200-204

Abnormal Behavior Detection of Escalator Passengers

Shifeng Wang[1]*
[1]*Anhui Huasu Co., Ltd, Chuzhou 232001, China*

Abstract: *This paper proposes an abnormal behavior detection algorithm for escalator passengers based on target detection in the context of escalators. The optimal prediction box is selected through cluster analysis. The behavioral features of passengers are enhanced by combining up-sampling and down-sampling methods, strengthening the feature fusion capability of the neural network. Based on the trained model, the behavior of escalator passengers is classified and detected, identifying abnormal behaviors. Experimental results show that this algorithm has a good detection effect on abnormal behaviors of escalator passengers and can effectively improve the safety of escalator passengers.*
Key words: *Escalator; YOLOv5; Abnormal behavior detection*

Date of Submission: 15-05-2026 Date of Acceptance: 31-05-2026

I. Introduction

With the process of national urbanization, the urban population is gradually increasing, especially in public places such as shopping malls, hospitals, and railway stations, where the flow of people is more dense. Escalators are widely used in public places like railway stations to provide convenience for people's travel. However, due to passengers' insufficient safety awareness and other unexpected factors, safety accidents on escalators frequently occur, posing great harm to people's lives and property. Therefore, research on identifying abnormal behaviors of passengers is beneficial for ensuring people's safety, preventing escalator safety accidents, and improving public safety.

In recent years, with the development of computer vision, many scholars at home and abroad have applied computer vision to the field of elevator detection and monitoring. Tian Lianfang et al. [1] identified abnormal behaviors of passengers through human skeleton sequences. Ji Xunsheng [2] applied the Tiny YOLOv3 algorithm to the detection of abnormal behaviors of passengers in escalator scenarios and improved its network detection structure. Nie Hao et al. [3] proposed a two-stream convolutional neural network video abnormal behavior detection algorithm for abnormal behaviors of people in videos.

Huang et al. [4] improved the Hopfield neural network to extract human skeletons of passengers to determine their behaviors. Nunez Marcos et al. [5] classified and detected passenger behaviors using convolutional neural networks. Thanh-Hai Tran et al. [6] extracted human features from data images using multimodal features from Kinect sensors for human behavior recognition.

This paper proposes an abnormal behavior detection algorithm for escalator passengers based on YOLOv5. Its algorithm model is small, has fast detection speed, and high accuracy, performing well in real-time detection and accuracy. It can quickly and accurately identify abnormal behaviors of passengers, perform real-time detection and classification, prevent major safety accidents, prevent further accidents, and ensure the safety of passengers.

II. Object Detection Algorithm

2.1 YOLO Algorithm

The YOLO algorithm is a single-stage object detection algorithm that uses only one convolutional neural network to predict objects of different categories. It defines object detection as a regression problem using an end-to-end detection method, which has the advantages of good detection effect and high detection speed.

2.2 YOLOv5 Principle

YOLOv5 is a detection algorithm proposed by Ultralytics based on YOLO. The convolutional neural network of YOLOv5 divides the input video image into S*S cells, and each cell is responsible for detecting objects whose center falls within itself. YOLOv5 uses the model to input the entire image into the network. At the input end, image preprocessing is performed through image slicing and splicing, feature sampling, and normalization. In the backbone network, the convolutional network is used to extract the feature structure of the

image. In the neck network, the diversity and robustness of the feature structure are enhanced. In the output network, the loss function is calculated, and non-maximum suppression (NMS) is used to filter out the best bounding box containing the detected object class and location information, completing the detection of the target object.

There are four versions of the YOLOv5 algorithm: YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x. This paper will focus on YOLOv5s, as other versions are based on this version with increased network depth, width, and number of convolutional kernels. The YOLOv5 detection algorithm can be divided into four modules: input end, backbone network, Neck network, and prediction output end.

The input end preprocesses the input image. The backbone network uses the Focus structure to slice and convolve the image to form feature maps. The CSP structure is used to increase the learning ability of CNN while maintaining accuracy with lightweight, reducing computational bottlenecks and memory costs. The NECK network combines FPN+PAN to use up-sampling and down-sampling methods to strengthen semantic and localization features and enhance the network's feature fusion capability, aggregating parameters for different detection layers and outputting prediction feature maps. The output end calculates the loss function and filters prediction boxes using weighted NMS to output prediction boxes and categories.

2.3 Prediction Box and Filtering

For each cell, B bounding boxes and their confidence levels are predicted. Confidence is divided into two aspects: the probability of an object being contained within the bounding box, denoted as Pr(object), and accuracy.

Pr(object) determines whether an object is within the bounding box. The accuracy of the predicted bounding box is represented by the Intersection Over Union (IOU) of the predicted box and the actual box.

$$IOU = \frac{P \cap A}{P \cup A} \tag{1}$$

Where P is the area of the predicted box, A is the area of the actual box. A larger IOU value indicates a higher overlap between the predicted box and the true box, and better prediction performance. Confidence is:

$$Confidence = Pr(Object) \times IOU_{pred} \tag{2}$$

The position and size of the predicted bounding box are represented by four values (x, y, w, h). (x, y) are the center coordinates of the predicted bounding box with the top-left corner of its cell as the origin. (w, h) are the width and height ratios of the bounding box to the entire image. All values are in the range [0, 1]. The confidence value of the predicted box is denoted by c. Therefore, each predicted box's prediction value contains 5 values: (x, y, w, h, c).

Assuming that the offset of each cell from the top-left corner of the cell as the origin is c_x, c_y, and the width and height of the predicted box are p_w, p_{h}, then the formula for the predicted box is:

$$b_x = s(x) + c_x \tag{3}$$
$$b_y = s(y) + c_y \tag{4}$$
$$w_b = w \times p_w \tag{5}$$
$$\mathrm{h}_b = \mathrm{h} \times p_{\mathrm{h}} \tag{6}$$
$$Pr(object) \times IOU(object, b) = s(c) \tag{7}$$

Where $s(x, y)$ is the coordinate of the initial prediction point relative to the grid with the top-left corner as the origin. Cluster analysis is performed on manually labeled bounding boxes to find the most suitable prediction box size. The key is to improve the IOU score, which depends on the size of the Box. The formula for finding the appropriate size is:

$$D(box, centroid) = 1 - IOU(box, centroid) \tag{8}$$

In the formula, centroid is the central bounding box selected during clustering, box is other bounding boxes, and D is the distance between the two. The larger the IOU, the closer the distance.

During the training and prediction process, multiple prediction boxes are generated. To select the best prediction box for actual object detection, a method called "Non-Maximum Suppression (NMS)" is applied to each class of objects to filter out the optimal detection box. First, a confidence threshold is set. All prediction

boxes for each class are compared, and prediction boxes with confidence below the threshold are discarded. The prediction box with the highest confidence in the batch is retained. Through continuous iteration, prediction boxes are filtered, inefficient prediction boxes are continuously removed, ensuring that one best prediction box is retained for each detected object.

2.4 Loss Function

The YOLO algorithm's loss function includes prediction box position loss, confidence loss, and classification loss. The position loss function is GIoULoss, which handles the problem of no overlap between the predicted box and the actual box by adding the smallest bounding box containing both. The formula is:

$$GIoU = IOU - \frac{A_c - A_u}{A_c} \tag{9}$$

Where A_c is the area of the smallest box containing both the predicted box (Box) and the actual box (ground truth box), and A_u is the area of the union of the predicted box and the actual box.

The confidence loss function is FocalLoss, which considers the severe imbalance between positive and negative samples in detection, achieved by increasing the loss function for difficult-to-classify categories. The formula is:

$$FL(p_t) = -(1-p_t)^{\gamma}\log(p_t) \tag{10}$$

The classification loss function is LogitsLoss. First, the predicted output undergoes a sigmoid transformation, and then the binary cross-entropy between the transformed result and the true value is calculated to obtain the loss of the predicted box class. The formula is:

$$Logits = -\frac{1}{n}\sum_{i=1}^{n}[y_i\log(x_i) + (1-y_i)\log(1-x_i)] \tag{11}$$

2.5 Algorithm Detection Process

The abnormal behavior detection process for escalator passengers is shown in Figure 2. Data is input into the algorithm, preprocessed, and then the data features of the image are enhanced to extract passenger behavior features and enhanced feature information to output prediction feature maps. Subsequently, the personnel status is analyzed, and behavior classification is performed through a classifier to output prediction results.

(1) Image Preprocessing: After the entire image is input through the input end, the image is uniformly scaled to a fixed network size. The image can set different width and height anchor boxes according to different datasets. Through manual labeling, the selected anchor box region image uses the Mosaic data augmentation method to crop and randomly splice 4 images, which enriches the data and improves training speed. Then, the image is input into the backbone network.

(2) Feature Preprocessing: In the backbone network, the image is first sliced. Sampling is performed every other pixel to concentrate and expand information, concentrating W and H information into the channel space. The input channels are expanded by 4 times, increasing the receptive field of each point and reducing the loss of original information. That is, the spliced image becomes a 12-channel mode compared to the original RGB three-channel mode. Subsequently, the new image obtained is further processed by convolution to extract features, finally obtaining a two-fold down-sampled feature map without information loss.

(3) Feature Extraction: In the backbone network, the FPN structure uses up-sampling to transmit and fuse feature information from top to bottom. Then, the PAN structure uses a feature pyramid from bottom to top to down-sample and transmit localization features. Through these two structures, feature aggregation is performed on different detection layers of different backbone layers to output prediction feature maps.

(4) Classification Prediction: Through training, the labeled behavior categories are analyzed, and the extracted behavior features are learned. Data features are reinforced through deep learning. Through the classification network, the feature model is continuously iterated and optimized to obtain the best classification model. The

loss function is calculated, and the best classification model after training iteration is used to classify the prediction results.

III. Experiment and Analysis

3.1 Dataset Production

To detect abnormal behaviors of passengers in escalator scenarios, video images were collected from actual escalator scenes at different times and light intensities. The video images were converted into image data using the OpenCV development library. Useless data was removed, and 6000 images were selected to create the dataset.

The dataset was divided into two groups: 70% for the training set and 30% for the test set. In the dataset, passenger behaviors were labeled and classified into categories such as up, down, bow, and shaking, representing standing, sitting (including falling, lying down), bending over, and shaking, respectively. Sitting and shaking were set as abnormal behaviors.

3.2 Experimental Results and Analysis

The optimal model, continuously iterated and trained on the training set, was selected to test the test set. The algorithm model detected human behaviors in the data, and the actual detection results are as follows:

The detection algorithm proposed in this paper achieved a recognition rate of 99.5% for abnormal behaviors in single-person scenarios and 97.6% for abnormal behaviors in multi-person scenarios on the test set, showing good detection performance.

Using critical criteria commonly applied in fall detection, two main indicators, sensitivity and specificity, were selected as evaluation indicators. Their calculation formulas are as follows:

$$Sensitivity = \frac{TP}{TP + FN} \tag{12}$$

$$Specificity = \frac{TN}{TN + FP} \tag{13}$$

In the formulas, TP is the number of times that falling behavior is detected in all test video data, and FN is the number of times that falling behavior is not detected in the video data. TN is the number of times that no falling behavior is detected in the video data that does not contain falling behavior, and FP is the number of times that falling behavior is detected in the video data that does not contain falling behavior.

Subsequently, experiments were conducted by different experimenters in single-person and multi-person environments. The first set of video data included various normal and abnormal behaviors in a single-person environment with different passengers and light intensities. The second set of environments included normal and abnormal behaviors of passengers in a multi-person environment with different light intensities. Through actual testing, it was found that when the escalator environment is single-person and has good lighting, the sensitivity and specificity of detection are relatively high, reaching 99.3% and 98.9%. In single-person situations with weak lighting, night, rain, and other conditions, the sensitivity and specificity of detection are 98.8% and 97.8%, showing excellent detection effects.

In multi-person scenarios with good lighting, the sensitivity and specificity of detection are 97.5% and 96.6%. In multi-person scenarios with weak lighting, the sensitivity and specificity of detection are 96.3% and 95.7%. The detection effect of the algorithm in multi-person scenarios with occlusion still needs further optimization.

IV. Conclusion

This paper proposes an abnormal behavior recognition algorithm for escalator passengers based on YOLOv5. By collecting video of escalator passengers' behaviors on-site, extracting their behavioral features, and training, abnormal behaviors of passengers are detected. Experimental results show that the abnormal behavior recognition algorithm proposed in this paper has a good detection effect on abnormal behaviors of escalator passengers. On the test set, the recognition rate for abnormal behaviors in single-person scenarios reached 99.5%, and for multi-person scenarios, it was 97.6%, demonstrating good detection performance.

From the experimental results, it can be seen that weak lighting, multi-person occlusion, and other conditions in the environment will affect the recognition effect to a certain extent. In multi-person and weak

lighting conditions, the recognition effect has certain shortcomings and there is room for improvement, which is also the direction for future research and optimization.

References

[1]. Tian L F, Wu Q C, Du Q L, et al. Abnormal behavior recognition of escalator passengers based on human skeleton sequence [J]. Journal of South China University of Technology, 2019, 47(04): 10-19.
[2]. Ji X S, Teng B. Escalator abnormal behavior detection based on deep neural network [J]. Laser & Optoelectronics Progress, 2020, 57(6): 140-149.
[3]. Nie H, Xiong X, Guo Y D, et al. Video abnormal behavior recognition algorithm based on deep learning [J]. Modern Electronic Technology, 2020, 43(24): 110.
[4]. HUANG X, HAO K, DING Y. Human fringe skeleton extraction by an improved Hopfield neural network with direction features [J]. Neurocomputing, 2012, 87(1): 99-110.
[5]. Nunez-Marcos, A, Azkune, G, Arganda-Carreras, I. Vision based fall detection with Convolution Neural Networks [J]. Wireless Communications and Mobile Computing, 2017(12): 1-16.
[6]. Thanh-Hai Tran, Thi-Lan Le, Van-Nam Hoang, et al. Continuous detection of human fall using multimodal features from Kinect sensors in scalable environment [J]. Computer Methods and Programs in Biomedicine, 2017, 146(1): 151-165.