IEICE TRANSACTIONS

on Information and Systems

DOI:10.1587/transinf.2024EDL8011

Publicized:2024/06/26

A PUBLICATION OF THE INFORMATION AND SYSTEMS SOCIETY

The Institute of Electronics, Information and Communication Engineers

Kikai-Shinko-Kaikan Bldg., 5-8, Shibakoen 3 chome, Minato-ku, TOKYO, 105-0011 JAPAN

LETTER

SH-YOLO: Small Target High Performance YOLO for abnormal behavior detection in escalator scene

Shuoyan LIU†, *Members*, Chao LI†, Yuxin LIU† and Yanqiu WANG††, *Nonmember*

SUMMARY Escalators are an indispensable facility in public places. While they can provide convenience to people, abnormal accidents can lead to serious consequences. Yolo is a function that detects human behavior in real time. However, the model exhibits low accuracy and a high miss rate for small targets. To this end, this paper proposes the Small Target High Performance YOLO (SH-YOLO) model to detect abnormal behavior in escalators. The SH-YOLO model first enhances the backbone network through attention mechanisms. Subsequently, a small target detection layer is incorporated in order to enhance detection of key points for small objects. Finally, the conv and the SPPF are replaced with a Region Dynamic Perception Depth Separable Conv (DR-DP-Conv) and Atrous Spatial Pyramid Pooling (ASPP), respectively. The experimental results demonstrate that the proposed model is capable of accurately and robustly detecting anomalies in the real-world escalator scene.

key words: *escalator, small target detection, attitude detection, YOLO*

1. Introduction

In public areas, escalators are an important mode of transportation for visitors. While they can bring convenience to people, abnormal accidents can lead to serious consequences. To this end, this paper attempts to use attitude detection algorithms to detect abnormal behaviors. Human attitude detection involves location and representation of human body parts (such as human bones) based on image, video, and other data [1]. Two types of algorithms based on Convolutional Neural Network (CNN) have been developed. The first is a two-stage object detection algorithm, represented by the Region-based Convolutional Neural Network [2] (R-CNN) and the Multi-Modal Pose Estimation algorithm. These algorithms generated candidate boxes [3] and then detected objects within them. They have high detection accuracy but are relatively slow. The second type is a single-stage object detection algorithm, represented by the You Only Look Once (YOLO) algorithm [4]. YOLOv8-Pose is the most advanced human attitude detection algorithm. Although this algorithm is fast, its accuracy has dropped significantly for small target. The size of the object on the escalator is small, which makes existing methods less applicable.

Therefore, this paper proposes the Small Target High Performance YOLO (SH-YOLO) model to detect abnormal behavior in escalator. Specifically, the SH-YOLO model first enhances the backbone network by introducing attention mechanisms. Subsequently, a small target detection layer is incorporated to enhance detection key points of small objects. Furthermore, the conv and SPPF are replaced with Region Dynamic Perception Depth Separable Conv (DR-DP-Conv) [5] and Atrous Spatial Pyramid Pooling (ASPP), respectively. Finally, the loss function is modified. The experimental results demonstrate that SH-YOLO model is capable of accurately detecting human behaviors in escalators. Although the detection speed is slightly slower, the frame rate has already reached 25 or more, which meets real-time detection requirements.

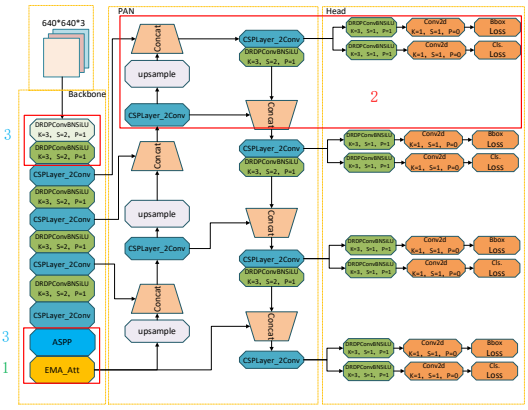

Fig.1 The network architecture of SH-YOLO model.

2.SH-YOLO: Small Target High Performance YOLO

SH-YOLO model is an extension of YOLO model, which has higher detection accuracy and timeliness performance compared to the original model, especially for small targets. We start by recalling the standard YOLO model, after which we propose the SH-YOLO model to our problem.

2.1 YOLO model

The YOLO model has existed numerous revisions. Different versions are proposed to adapt different scenarios. The fundamental structures of these models are essentially identical. The YOLO v8 model is a widely used tool in video surveillance. The model comprises four principal components: the input, the backbone network, the neck, and the prediction layer (head). The backbone networks are C2f and SPPF. The neck layer and the head layer are designated as the PANet and Decoupled Head, respectively.

†The author is with China Academy of Railway Sciences Corporation Limited, 180-8585 China.

††The author is with School of Electronic Information Engineering,Beijing Jiaotong University, Beijing,China.
13052777873@163.com

2.2 Model Improvement

This paper proposes the SH-YOLO model and its architecture is shown in the Figure 1, where the parameters K, S, P represent the size of the convolutional kernel, the step size, and the size of the boundary expansion of the feature map to be convolved, respectively. In addition, in the Head module, Bbox Loss represents regression box loss, while Cls Loss represents classification loss. All the improvements from the original version are highlighted in red boxes.

2.2.1 EMA attention mechanisms

The self-attention mechanism has been widely employed in various tasks. It calculates the representation of each position through establishing long-range relationships. However, its computational cost is considerable. To this end, this paper presents an expectation maximization approach to formulate the attention mechanism and iteratively estimate a more compact set of foundations. The attention graph is computed through basic weighted summations, resulting in a low-rank representation. The EMA module is robust to input variance and hence has efficient memory and computational requirements. Figure 2 depicts the network structure. In this module, the size of the network's feature map and the number of channels remain constant.

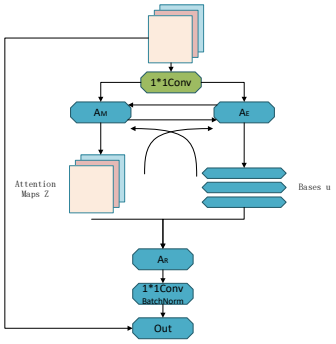

Fig.2 EMA structure.

2.2.2 Small target detection layer

Though YOLOv8 employs the multi-scale fusion to improve accuracy in small object detection, human joints with smaller pixels may still cause missed detection. The main reason is that YOLOv8 has a large downsampling factor. The downsampling rates are 8 times, 16 times, and 32 times. Therefore, SH-YOLO model adds the small object detection layer (marked as 2 in the red box in Figure 1). It takes 4 times downsampling from the backbone network as input, and after passing through the neck and prediction layers. and deep feature maps after concatenation.

If we consider an input image of 640 × 640 as an example, the output feature map of SH-YOLO has a resolution of 160 × 160, 80 × 80, 40 × 40, and 20 × 20, whereas the output feature maps of YOLOv8 have resolutions of 80 × 80, 40 × 40, and 20 × 20.

Adding a small object detection layer can enhance the network's ability to detect small objects and improve detection efficiency. Nevertheless, it decreases the inference detection speed. The calculation formula for the parameter quantity of convolutional neural networks before and after improvement is shown in formula (1):

$$Parameters = K^2 \times C_i \times C_o + C_o \tag{1}$$

where C_i is the number of channels in the input image, C_o is the number of channels for outputting the image. Although the detection speed is slightly slower, the frame rate has already reached 25 or more, which meets real-time detection requirements.

The input channels of the improved four scale detection modules are 128, 256, 512, and 512, respectively. List of output feature maps (1 × 66 × 20 × 20, 1 × 66 × 40 × 40, 1 × 66 × 80 × 80, 1 × 66 × 160 × 160).

2.2.3 DR-DP-Conv and ASPP

The conv and SPPF is replaced with DR-DP-Conv and ASPP [6], respectively. Adding more filters to standard convolutional layers helps extract more visual elements, but this can be expensive. The proposed DR-DP-Conv approach uses learnable instructors in spatial dimensions instead of adding filters. This improves representation ability and keeps costs and translation invariance similar to standard convolutions.

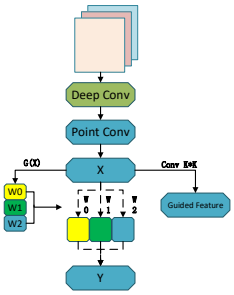

Fig.3 DR-DP-Conv structure.

Figure 3 shows the structure of DR-DP-Conv. First, the input is processed using depth-separable convolution to generate guide features. Then, the spatial dimension is divided into regions based on the guide features, each represented by a different color. In each shared region, the filter generator module generates filters for two-dimensional convolution. The number of parameters in the convolution module has been reduced after the improvement of depthwise separable convolution.

For the SPPF, the SH-YOLO model uses the ASPP module. It uses multiple parallel hole convolutional layers with different sampling rates, which are processed separately and then combined to create the final result. In figure 4, ASPP uses Atrous/Dilated Convolutions [7]. This

is very complex that use a larger convolutional kernel for pooling to achieve a larger receptive field. To address this, an empty convolution can be utilized, which permits the observation of more data without compromising resolution. The feature map of this module is 20 x 20 and has 1024 channels.

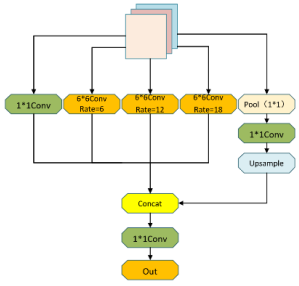

Fig.4 ASPP structure.

2.2.4 MPDIou

Existing loss functions cannot be optimized when the predicted and true bounding boxes have the same aspect ratio but different width and height. A new metric called MPDIoU [8] is used to explore the geometric features of horizontal rectangles. The formula is shown in Table 1.

Table 1 The formula of MPDIou

Algorithm: Intersection over Union with Minimum Points Distance

Input: Two arbitrary convex shapes: A, B $\subseteq S \in Rn$, width and height of input image: w, h

Output: MPDIou

1. For A and B,(x_1^A, y_1^A), (x_2^A, y_2^A)denote the top-left and bottom-right point coordinates of A, (x_1^B, y_1^B), (x_2^B, y_2^B)denote the top-left and bottom-right point coordinates of B.
2. $d_1^2 = (x_1^B, x_1^A)^2 + (y_1^B, y_1^A)^2$
3. $d_2^2 = (x_2^B, x_2^A)^2 + (y_2^B, y_2^A)^2$

4.MPDIou = $\frac{A \cap B}{A \cup B} - \frac{d_1^2}{w^2+h^2} - \frac{d_2^2}{w^2+h^2}$

2.2.5 Abnormal behavior detection in escalator scene

There are four of the most common abnormal behaviors in elevator scenarios, such as: the escalator pedestrians falling (PF), escalator pedestrians walking in the opposite direction (PW), escalator pedestrians running (PR) and escalator pedestrians squatting (PS). We attempt to detect them according to the following formula:

$$\begin{aligned} &\mathrm{PF}: \frac{1}{7}\sum_{j=1}^{7} y_j \geq \frac{1}{17}\sum_{i=1}^{17} y_i \\ &\mathrm{PW}: \vec{Y} \times \frac{1}{17}\sum_{i=1}^{17} \vec{Y}_i < 0 \\ &\mathrm{PR}: \vec{Y} > \alpha \frac{1}{17}\sum_{i=1}^{17} \vec{Y}_i \\ &\mathrm{PS}: y_{hip\ bone} \geq y_{knee} \end{aligned} \tag{2}$$

where y represents the vertical coordinate of a person's certain joint, $\vec{Y}$ represents the travel speed of a person, and α is the set threshold, $y_{hip\ bone}$ refers the vertical coordinate of a person's hip bone, and another y_{knee} refers the vertical coordinate of a person's knee.

3.Experiments and Results Analysis

In this section, we introduce a new dataset to evaluation of the effectiveness of our method and compared its performance with state-of-the-art methods on VOC. We construct a new escalator dataset from railway station (ED-RS), which is a collection of several passenger behavior in escalator. The ED-RS dataset contains 30924 pictures from 65 stations classified in the four categories, i.e. the escalator pedestrians falling(PF), escalator pedestrians walking in the opposite direction(PW), escalator pedestrians running(PR), escalator pedestrians squatting(PS), etc.

The environmental configuration used in this experiment is: (1) CPU is Intel (R) Core (TM) - i7-8750H; (2) GPU is NVIDIA GeForce - RTX 3080; (3) Memory is 16GB; (4) Operating system is Win10. During the experiment, YOLOv8-Pose code served as the baseline, employing SGD optimization for training. The training parameters include an initial learning rate of 0.01, periodic learning rate decay of 0.2, weight decay of 0.0005, and a batch size of 8. The model is trained for 200 epochs. Additionally, pre-learning and regular momentum are set to 0.8 and 0.937, respectively. The input image size used for the experiment was 640×640.

We first compare the performance of the proposed approach with existing model on VOC dataset, such as CASANet, OpenPose, and Mask RCNN. Mean Average Precision (mAP), and Latency are used as performance evaluation indicators. Table 2 illustrates the SH-YOLO model outperforms the other methods in detecting objects. The main reason is that the implementation of a multi-downsampling methodology significantly enhances the capability to extract intricate and profound features from the input data. Furthermore, the attention mechanisms facilitate a focused attention on discriminative representations in each level.

Table 2 Comparison of different algorithms in the VOC dataset

Model	mAP50(%)	mAP50-95(%)	Latency/ms
Mask-RCNN	47.9	21.5	11.9
OpenPose	76.9	51.8	15.1
CASANet	78.8	54.1	14.5
SH-YOLO	80.3	57.6	31.9

Subsequently, we verify the effectiveness of the SH-YOLO model in detecting small target using ED-RS dataset in the table 4. It can be seen that the SH-YOLO model has greatly improved accuracy compared to the YOLO v8 model. Although the inference speed has shown a downward trend, it still meets the task of real-time detection. Figure 5 shows the key point detection results using YOLO and SH-YOLO inference in the same scenario. Compared with the YOLO

v8 model, the SH-YOLO model has better recognition accuracy.

Table 3 Research Results of Network Overlay Optimization in ED-RS

method	mAP50 (%)	mAP50-95 (%)	Latency (MS)	Parameters (M)
YOLOv8	68.9	47.6	22.2	43.7
SH-YOLO	75.2	50.4	35.9	52.6

（a） YOLO v8 model（b） SH-YOLO model

Fig.5 Comparison of results of YOLO v8 and SH-YOLO models.

We investigate how the detection performance is affected by each module. In order to ensure the reliability and consistency of the results, we adopt the same experimental environment and evaluation criteria. Table 3 illustrates that the network superposition experiment provides an intuitive representation of the impact of each improvement point on the speed and accuracy of the model. First, adding a small object detection layer improves detection accuracy for small objects. However, the speed decreases due to an increase in the quantity of parameters. The main reason is that the introduction of a small number of new structures resulted in an increase in the number of parameters and a decrease in inference speed of 1.5ms. Furthermore, the attention mechanism makes the accuracy of the model mAP@.5 and mAP@.5 :. 95 increased by 1.9% and 4.9%, respectively. Finally, all Convs are replaced with DR-DP-Conv, mAP@.5 and mAP@.5 :. 95 increased by 0.78% and 0.56% respectively. Replacing the original model's CIou with MPDIou improves accuracy during training, but does not affect speed during testing.

Table 3 Research Results of Network Overlay Optimization in VOC

(method)Model	mAP50 (%)	mAP50-95 (%)	Latency (MS)	Parameters (M)
YOLOv8-l	77.3	53.6	23.6	43.7
(Replace)DR-DP-Conv	77.9	53.9	24.5	43.9
(Replace)ASPP	78.1	54.1	25.2	43.9
(Replace)MPDIou	78.2	54.2	25.2	43.9
(Add)EMA-Attention	79.7	56.9	26.7	44.0
Small target detection layer	80.3	57.6	31.9	52.6

Detection of abnormal events in an escalator is of great importance to the station management and to the public safety, as it brakes untimely, the abnormal events of which may lead to fatal consequences further. Table 4 tests the SH-YOLO model using ten videos collected real-time from the Lanzhou and Beijingchaoyang railway station, respectively. Looking closely at the Table 4 from the railway station perspective, we can find out that the accuracies are generally higher in the BeijingChaoyang station than in the Lanzhou Station. We believe the difference is related to the layout of the video from these two stations. The analysis of Table 4 by abnormal categories is follows: the best performance is achieved to detect the escalator pedestrians walking in the opposite direction. The main reason is that the abnormal behaviors relatively simple and highly discriminative. It is usually difficult to accurately detect escalator pedestrians squatting. Because this behavior is easily occluded. Also, due to the lack of enough illumination, many behaviors are not accurately estimate.

Table 4 The performance of the proposed model on ED-RS dataset

	PF(%)	PW(%)	PR(%)	PS(%)
STA.BeijingChaoyang	89.2	92.6	91.4	82.7
STA. Lanzhou	86.5	93.4	90.8	83.5

4.Conclusion

This paper proposes an improved SH-YOLO model to detect abnormal behavior in the escalator scene. The traditional YOLO model cannot well detect the small target objects. In view of this, we solve this problem by introducing attention mechanisms, adding a small object detection layer, and improving the backbone network. In future work, we plan to incorporate other cues into consideration to improve the detection performance.

Acknowledgements

This work was partially support by the China State Railway Group CO., Ltd (N2023X005) and the fund of China Academy of Railway Sciences (2022YJ073).

References

[1] Z. YIN, X. WANG, L. LI. “Optimization of Human Body Attitude Detection Based on Mask RCNN”. Proc. Int. Conf. on Orange Technology, pp.18-21, 2020.

[2] R. Girshick, J. Donahue, T. Darrell, et al. “Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation”. Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 580-587, 2014.

[3] Q. Liu, Z. Zhao, J. Wang, et al. “High Performance YOLOv5: Research on High Performance Object Detection Algorithms for Embedded Platforms” Journal of Electronics and Information Science, vol.45, no.6, pp. 2205-2215, 2023.

[4] J. Redmon, S. Divvala, R. Girshick, et al. “You Only Look Once: Unified, Real-Time Object Detection”. Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 770-788, 2016.

[5] J. Chen, X. Wang, et al. “Dynamic Region-Aware Convolution.” Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR), pp.8060-8069, 2021.

[6] D. XIAO, M. WANG, L. ZHAO, et al. “Dual ASPP for Lightweight Semantic Segmentation on High-Resolution Image.” Proc. International Symposium on Computational Intelligence and Industrial Applications (ISCIIA), pp.1-6,2020.

[7] F. Yu, V. Koltun. “Multi-Scale Context Aggregation by Dilated Convolutions”. Proc. IEEE Int. Conf. on Learning Representations, pp. 1-13, 2015.

[8] J. CHENG, J. YUAN, X. HU, et al. “Lightweight model of remote sensing ship classification based on YOLOv7-tiny improvement.” Journal of Physics: Conference Series, vol.26,no.1,pp.2-6,2023.