Detection of public chair uncivilized lying behavior based on YOLOv5

Liu-Ying Zhou[1], Dong Wei[*1, 2], Yi-Bing Ran[3], Chen-Xi Liu[4], Si-Yue Fu[5], Zhi-Yi Ren [6]

[1] Department of Electrical and Information Engineering, Beijing University of Architecture & Beijing Key Laboratory of Intelligent Processing for Building Big Data, Beijing
E-mail: 1522588620@qq.com
[*1, 2] Same as first author
E-mail: weidong@bucea.edu.cn
[3] Siemens Ltd., China, Beijing
E-mail: rybing2009@163.com
[4,5,6] Same as the first author
E-mail: 2108110021025@stu.bucea.edu.cn; fusivue0718@163.com; 1052190579@qq.com

Abstract. To monitor and manage the uncivilized lying behavior of the urban public chair, this paper studies the urban public chair lying behavior image recognition algorithm based on YOLOv5. At present, the object studied in the field of object detection is a single object, but for the uncivilized lying behavior of the urban public chair, the object to be measured is composed of the chair and the human posture combined, so its characteristics will be more complex. To accurately identify the uncivilized lying behavior of urban public chairs, this paper designs various dataset classification schemes according to the feature differences among various combined objects, and tests them using the YOLOv5x object detection algorithm. The experimental results show that the best performance index mAP is 97.38%, and the overall average recall rate (Recall) is 95.51%, which can effectively realize the detection and identification of uncivilized lying public chair behavior.

Keywords: uncivilized lying behavior of urban public chairs; deep learning; object detection; YOLOv5x; image processing

1. INTRODUCTION

In recent years, with the development of the economy and the improvement of science and technology levels, China's urbanization scale has also achieved a huge expansion. But uncivilized behaviors such as lying down in public chairs have seriously damaged the image of the city. To help the relevant city departments to realize the standardized management of uncivilized behavior, this paper designs an object detection model to accurately detect and identify the reclining public chair behavior appearing in images. When the traditional object detection method is used to detect the input images, the pre-set sliding window is used to extract the candidate box, then the features of each candidate box are extracted, and finally, the classifier is used to identify them. For example, Haar + Adaboost[1], DPM [2], and Hog + SVM[3] are all relatively classical object detection algorithms. However, such traditional object detection algorithms have the disadvantages of poor robustness and high computational cost[4].

Under the continuous exploration of many scholars at home and abroad, the field of object detection begins to develop rapidly, and high-performance algorithms emerge one after another. Nowadays, in the field of object detection, the used algorithm architecture is divided into two types: One-stage and Two-stage. Two-stage decomposes the whole object detection process into two parts. The first step is to determine the approximate location of the object and generate a candidate box, and the second step is to classify the objects in each candidate box to correct the position of the candidate box, to determine the final prediction box of the object through a given threshold. Representative algorithms for the Two-stage type include the RCNN series, etc, which have the advantages of high detection accuracy but also have the disadvantages of large computational quantity and slow detection speed. One stage is the overall regression problem of treating the whole process of object detection as the category probability that the object belongs to be detected, and the position coordinate value of the image where it is located. Its advantage is the fast detection speed. However, without generating candidate boxes generation process, One-stage-type algorithms generally have lower detection accuracy than Two-stage, including YOLO series algorithms, SSD series algorithms, Efficient Det, and so on.

In this paper, the YOLOv5 algorithm was used to identify the reclining public chair phenomenon. Currently, the YOLOv5 series has derived a variety of different algorithm models. The overall network structure of these algorithms is roughly the same, but there are subtle differences in the network depth and width, so the number of parameters included in each version is also different. The more parameters, the detection accuracy is generally relatively high, but the detection speed will be correspondingly slow. To balance the relationship between the detection accuracy and the detection speed, and to try to improve the accuracy based on real-time detection, the application of YOLOv5x is selected as the algorithm model of the experiment. The YOLOv5 algorithm has both a good detection rate and a short training time, which has become the priority choice of the current object detection network.

Wang et al.[5] proposed an underwater object detection algorithm based on attention improvement YOLOv5, introducing a two-stage cascade of CSP structure and visual channel attention mechanism module SE in the original backbone network, which effectively improves the detection accuracy of the model for underwater objects. Aveiro et al. [6] have proposed using YOLOv5 to detect the GW signal of BNS merging in the raw interferometer data of existing detectors, which showed that YOLOv5 can be used to infer physical primal parameters in real-time when

integrated into more complex pipelines. Keles et al.[7] studied the test results of two algorithmic models of YOLOv5 and YOLOX on the VisDrone2019Det dataset. According to the results, the YOLOv5 algorithm performs better on the small object detection problem when the other processing mechanisms are consistent. Huang et al.[8] proposed a YOLOv5-based bone marrow cell detection algorithm, which is trained by minimizing a new loss function, and effectively reduces the difficulty of bone marrow cell detection due to the dense and diverse cell distribution. Raian et al.[9] proposed a method to use YOLOv5 to classify and locate vehicles from densely crowded images. Zhu et al.[10] proposed the TPH-YOLOv5 model for UAV(Unmanned Aerial Vehicle) capture scene object detection. With YOLOv5 as the basic framework, the model replaces the original predictor head with the predictor head of the transformer predictor head (TPH) and adds the convolutional block attention model (CBAM) and another predictor's head, effectively enhancing the model's ability to extract the object feature information at different scales. Gu et al.[11] proposed a real-time vapor sensing system that uses the YOLOv5 model as the basic framework and incorporates data augmentation, Transformer, Bag-of-Freebies, etc. Reddy et al.[12] have proposed a hand-drawn circuit identification model, equipped with YOLOv5 as a circuit component detection model, and incorporating a node identification method based on the Hough transform.

This paper first conducts data classification analysis and data preprocessing of image data and designs various dataset classification schemes. Then, by training and testing different dataset schemes by using the YOLOv5x algorithm, a dataset design scheme and algorithm model that can effectively help accurately identify the detection of urban uncivilized public chair reclining behavior are obtained.

2. RELATED WORK

2.1 Data analysis and preprocessing

At present, in the field of object detection, the detected objects in the used dataset are mostly composed of single objects, such as vehicles, helmets, masks, etc. Because the object detection algorithms basically adopt supervised learning methods, they need to manually classify and annotate the pending objects of the dataset in advance. However, this kind of object composed of a single object has similar characteristics to the same type of object. Even in the case of multi-object detection, all types of objects are relatively independent. Therefore, it is easier to classify and annotate them, and it is less difficult to process the dataset. However, for the behavior detection of reclining public chairs proposed in this paper, the object to be measured is composed of a chair and human objects. Moreover, because there are differences in the types of chairs, and human lying posture is also different, there are large or small differences between the objects formed after the combination of the two objects. To this end, the pre-processing of data is extremely important.

As shown in Figure 1, the types of public chairs in life can be roughly divided into three kinds, namely, back with armrest, back without armrest, and bench. This paper studies the following four uncivilized lying postures (take the armrest and the backrest type public chair as an example): 1) lying on the side; 2) lying on your back; 3) lying prone; 4) sitting sideways.



(a) chairs type (b) uncivilized postures type

Fig.1 public chair standard and character posture analysis

The large differences between the various types of public chairs, and the human body's uncivilized lying postures are also quite different, and the types of objects to be tested composed of these two objects have a more complex diversity. In order to better classify and annotate each combined object, the research team conducted a detailed hierarchical classification of each type of sample data collected, and analyzed it in detail, as shown in Figure2.

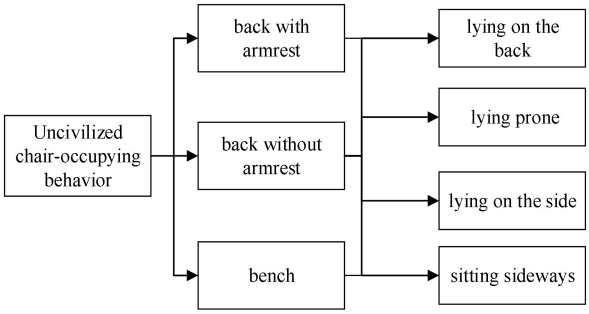


Fig.2 The classification form of the object to be tested

According to the analysis, because of the different types of public chairs, the two types of chairs, back without armrest and back with armrest, are basically the same in general shape, with only the difference between the left and right sides of the chair with or without armrest. The difference between the two chairs in the overall characteristics is small, so we put the two into the same category. For the bench type, there is an obvious difference between the first two chairs in shape, so it is regarded as a single category alone. As far as human postures are concerned, this paper is not as detailed in the detection of human postures as face detection, so whether it is lying on the side, lying on the back, or lying prone, these three postures are generally fewer different, so these three postures are classified as one category. The posture of sitting sideways is significantly different from other postures, so it is classified into one category alone.

In addition, because lying public chair behavior occurs outdoor, some object features in the different outdoor scenes may error phenomenon, in order to reduce the impact, this paper adopts Copy-Paste [13] method for data enhancement, by randomly pasting some large object images into outdoor background images, not only provides enough background negative samples for algorithm model training but also solve the problem of the insufficient number of small and medium object samples, improve the model generalization ability, and the detection accuracy of small objects.

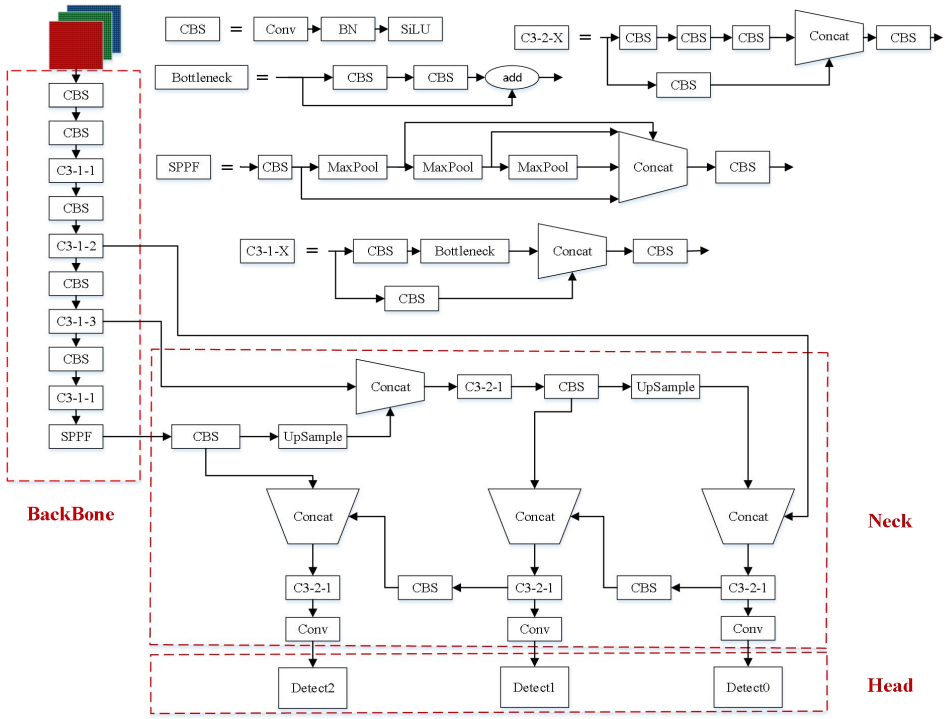


Fig.4 YOLOv5 model network structure

After the model is trained, the preliminary test results of the dataset designed by the above method show that there is more serious misdetection phenomenon for the normal and civilized sitting posture. The reason for the research team's analysis of this phenomena is that when the posture of sitting sideways is annotated, the upper body characteristics of humans are inevitably included in the marking box, which overlaps with the object characteristics of civilized sitting posture to a large extent, and thus leads to the occurrence of this misdetection phenomenon. Therefore, in addition to improving the annotation method, this paper also selects the images of some civilized sitting postures as the negative samples in the algorithm model according to the proportion of 10% of the overall positive sample data. Considering the above analysis, by analyzing and sorting out the existing data, the final dataset design scheme is obtained, as shown in Figure 3.

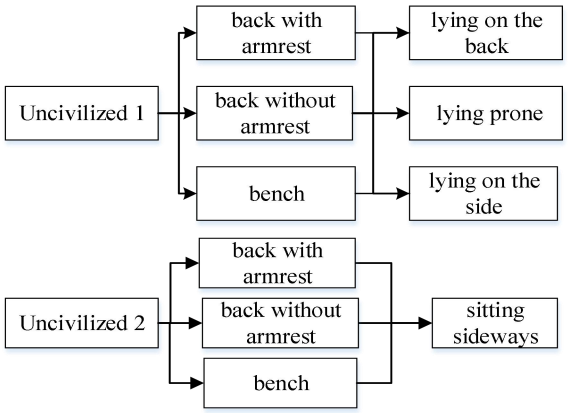


Fig.3 Dataset classification scheme one: uncivilized1 and uncivilized2 indicate the categories of the corresponding object

2.2 Algorithmic model design

2.2.1 YOLOv5 Network Structure

The network structure of the YOLOv5 series object detection algorithm based on One-Stage architecture can be roughly divided into four parts: input terminal, Backbone network layer, Neck network layer, and Head terminal. images have entered the model through the input terminal and features are extracted in the Backbone network layer. Then the extracted features are further multi-scale feature fusion in the Neck network layer. Finally, It will perform a series of loss calculations and determine the predictions in the Head layer. The overall network structure of the YOLOv5x(v6.1) is shown in Figure 4.

2.2.2 The Backbone Network Layer

The Backbone network layer mainly includes the CBS (Convolution BatchNormalization SiIU) structure, the C3 (CSP Bottleneck with Three Convolutions) structure, and the SPPF (Spatial Pyramid Pooling-Fast) structure. The residual modules are integrated into the C3 structure, which can adjust the depth of the backbone network by adjusting the depth of the residual module. It is worth mentioning that, before version v6.0, the first layer of the Backbone network used the Focus module. The Focus module takes the adjacent pixels of each 2 * 2 of the input image as a separate patch and then combines the pixels in the same position in each patch into a feature map, after which a 3*3 convolution layer is used for feature extraction. These steps can be equivalently replaced by directly using a convolutional layer of 6*6 size, and the processing speed of feature extraction is faster than the former, making the Backbone network improve the processing speed of feature extraction of input images to a certain extent. In addition, before version v6.0, the last layer network structure of the Backbone network used the SPP (Spatial Pyramid Pooling)[14] structure rather than the SPPF structure.

As shown in FIG.5, the SPP structure processes the feature map output by the upper layer network in parallel through multiple MaxPools of different sizes, and then further integrates the respective output feature map, which is helpful in the multiscale problem of the object. The SPPF structure changes the multiple MaxPools from parallel to serial. By adjusting the size of the MaxPool, the output of the upper layer network has the same output result as the SPP structure after the SPPF structure processing, but the feature processing speed of the SPPF structure is faster than that of the SPP structure, which can improve the processing speed of the Backbone network.

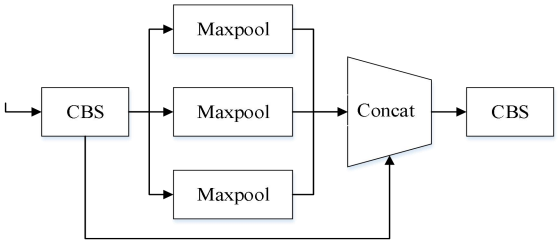


Fig.5 the SPP structure

2.2.3 The Neck Network Layer

In the Neck network layer part, the YOLOv5 algorithm adopts the fusion structure of the FPN[15] and the PAN[16], as shown in Figure 6.

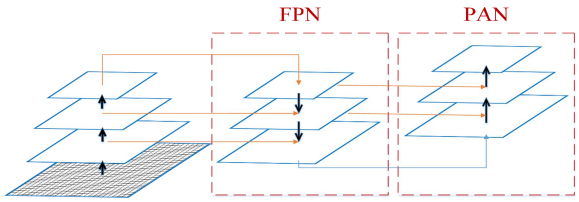


Fig.6 The FPN-PAN network structure

The FPN structure is a classical and effective enhanced feature extraction network, using a top-down feature pyramid structure. Multiple effective feature layers obtained from the backbone feature extraction network are multi-scale feature fusion in the FPN structure, which can effectively strengthen the semantic information of the object features and improve the detection performance. PAN structure is a bottom-up feature pyramid structure proposed by the PANet network. It transmits semantic information from low latitudes to high latitudes, which can effectively enhance the positioning information of objects and improve the detection performance of the model for small objects.

Fusion generating more discriminative multiscale features is a key to improving object detection performance [17]. As a feature fusion network widely used in the field of object detection, the feature pyramid FPN has a simple network structure and good performance. By constructing a top-down and lateral connection hierarchy, FPN transmits the high-level feature information extracted by the trunk classification network down, so as to supplement the semantic information of the lower level, to obtain high-resolution and strong semantic feature information, which can achieve good results on the detection of small objects.

Since the launch of FPN, many scholars have applied it in their own research, so a number of FPN-based feature fusion networks were born. The Chinese University of Hong Kong and Tencent Utu jointly proposed PANet, PANet is a neural network architecture for instance segmentation. PANet adds a bottom-up feature extraction path on the basis of FPN, which integrates the features of adjacent scales, accelerates the circulation of feature information, obtains the ability to accurately retain spatial information, and can properly locate the pixels to form a mask[18]. Pang et al. [19] proposed Libra R-CNN, based on FPN and PAN, achieves balanced learning by using balanced semantic features with the same and deep integration. Tan et al.[20] proposed a bi-directional weighted feature pyramid network structure (BiFPN), on the basis of FPN structure, the introduction of weighted strategy, through the training process by learning in adjusting the weight parameters, to learn the important semantic information in different input features, and through the FPN and PAN structure repeatedly multi-scale feature fusion, effectively improve the detection performance. Liu et al.[21] proposed an adaptive multiscale feature fusion network (AMFFN), which learns a fusion coefficient for features at each scale to achieve the adaptive fusion of features at multiple scales. Wang et al.[22] proposed the scale equilibrium pyramid convolutional network (SEPC). The author uses the internal correlation between different feature levels and introduces different weight parameters according to the richness of feature semantic information in various levels to realize the balanced fusion of feature information in different levels.

The YOLOv5x algorithm uses the classical FPN and PAN fusion structure, and it incorporates the C3 structure based on the conventional PAN. Because the residual module is integrated into the C3 structure, it can better alleviate the possible gradient disappearance problem during the model training process and can improve the computational efficiency to a certain extent. The FPN layer is responsible for conveying strong semantic feature information from the top down, while the PAN layer transmits strong localization feature information from the bottom up. The combination of each other integrates the parameters from different backbone network layers, which enhances the fusion ability of the network to the features and improves the detection accuracy of the model.

2.2.4 Head output layer

The input images are features extracted by the Backbone layer and the multi-scale feature fusion of the Neck layer, and the final formed feature map will be sent to the Head layer for the final border regression positioning and classification. In order to accurately determine the category and location of the object, the appropriate loss function must be used to complete the parameter update after each iteration in the head layer.

The calculated losses in the YOLOv5x algorithm are mainly divided into three parts: Classes Loss (Cla_Loss), Objectness Loss (Obj_Loss), and Regression Loss (Reg_Loss), which are expressed as follows:

$$\text{Loss} = a * Cla_Loss + b * Obj_Loss + c * Reg_Loss \quad (1)$$

The a, b, and c in the expression are all the weight parameters.

The loss function used by Classes Loss and Objectness Loss is BCEWithLogitsLoss and whose expression is:

$$BCEWithLogitsLoss = Sigmoid + BCELoss \quad (2)$$

$$Sigmoid(x) = \sigma(x) = \frac{1}{1 + exp(-x)} \quad (3)$$

$$L_{BCE}(x, y) = \{l_1, \ldots, l_N\}^{\top},$$
$$l_n = -w_n[y_n \cdot log\sigma(x_n) + (1 - y_n) \cdot log(1 - \sigma(x_n))] \quad (4)$$

The loss function used in the Regression Loss part is CIOU Loss (LCIOU), which introduces the minimum external rectangle of the prediction box and the real box on the basis of the IOU, and is calculated as follows:

$$L_{CIOU} = 1 - CIOU \quad (5)$$

$$CIOU = IOU - \frac{Distance(rho)^2}{Distance(C)^2} - \frac{v^2}{(1 - IOU) + v} \quad (6)$$

In the expression, rho is the Euclidean distance between the Ground Truth box and the central point of the prediction box, and C is the length of the diagonal line including both the Cround Truth box and the minimal closure region of the prediction box. Using CIOU Loss as a loss function of the regression part of the border not only well reflects the area of the overlapping area between the prediction box and the overlapping box but also calculates the proportion of non-overlapping regions. Therefore, the GIOU loss function better reflects the degree of overlap between the real box and the prediction box.

3. EXPERIMENTAL RESEARCH

3.1 EXPERIMENTAL PREPARATION

This experiment used GeForce RTX 3060 with 12GB for training and testing, and the batch size of the dataset was set to 4, and the number of iterations was 100 epoch. This paper uses the transfer learning method, using the pre-training weight yolov5x.pt and the hyperparameter file is hyp.scratch-high.yaml.

The multiscale training mode and the Mosaic data enhancement[23] method are opened, and the image input size is 640*640. The confidence threshold and intersection ratio thresholds were set to 0.3 and 0.5, respectively, and the remaining parameters adopt the default parameters set in the YOLOV5x model.

3.2 Experimental Methods

To verify the effectiveness of the above dataset classification scheme one, this paper designs dataset classification scheme two for comparison. In dataset classification scheme two, all the corresponding uncivilized positions in the different chairs are uniformly classified simply according to the type of public chairs. The classification situation is shown in Figure 7.

The images used to design dataset classification scheme two are exactly consistent with dataset classification scheme one, and the image annotation form is identical for both schemes.

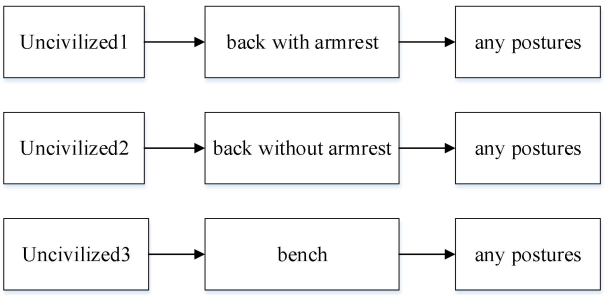


Fig.7 Dataset classification scheme two: uncivilized represent the category of the corresponding object

3.3 Experimental results

The dataset used for this experiment included 5878 images, including 3726 in the training set, 504 in the validation set, and 1645 in the test set. In this paper, Average Precision (AP), Recall, the harmonic mean of accuracy and recall(F1), and mean Average Precision(mAP) are used as the performance evaluation indicators of the experimental results, and the calculation method is PASCAL VOC2012 standard.

The calculation results of the performance indicators (mean values of all categories) of the two dataset classification schemes on the test set are shown in Table 1.

The speed in the table is the average detection time per image (speed = pre-process + inference + NMS).

Table 1 Test results for two dataset schemes

Id	Precision	AP	Recall	F1	mAP	Speed
One	97.73%	98.04%	95.46%	96.50%	97.38%	34.4ms
Two	91.24%	90.53%	89.59%	90.00%	90.53%	34.4ms

Comparing the detection results of the two dataset design schemes, the overall mAP of dataset scheme one is 97.38% and the overall Recall is 95.46%. However, the overall mAP of dataset scheme two was 90.53%, and the overall Recall was 89.59%. Compared with the two, the mAP of dataset scheme one has increased by about 7% and the Recall increased by about 6%. Obviously, the first dataset classification scheme is more suitable for the detection of urban uncivilized lying behavior.

3.4 Contrast it with the other algorithms

To further validate the effectiveness of this method, we compare other different object detection algorithm models against YOLOv5x on dataset classification scheme one. On the home-made dataset, we compared three object detection algorithms, YOLOv3-SPP, YOLOv5s, and YOLOv5x6, and the test results of the three models are shown in Table 2. Among them, the YOLOv3-SPP adds the SPP structure to the FPN structure of the Neck layer, which can better promote the feature fusion between different scales and help to improve the model detection performance. Compared with YOLOv5x, the YOLOv5s and YOLOv5x6 algorithms are only different in the depth and width of the network, and the overall structure is not much different from that of YOLOv5x.

Table 2 Test results of each model

Model	Precision	AP	Recall	F1	mAP	Speed
v3-SPP	97.47%	95.61%	75.47%	85.00%	95.61%	25.1ms
v5s	95.75%	96.27%	90.43%	93.00%	96.27%	9.4ms
v5x	97.73%	98.04%	95.46%	96.50%	97.38%	34.4ms
v5x6	96.70%	97.23%	96.84%	97.00%	97.23%	39.9ms

After comparing the test results of each model, we can see that YOLOv3-SPP and YOLOv5s have lower mAP and Recall, while YOLOv5x6 is detected more slowly compared to YOLOv5x. Despite the detection rate of the YOLOv5s is very rapid, YOLOv5 has met the real-time requirements of object detection. Therefore, the YOLOv5x model proposed here has both good accuracy and speed, and is more suitable for the detection of uncivilized lying behavior. Finally, the detection effect of the uncivilized lying behavior in the urban public chair by using YOLOv5x on the dataset classification scheme one is shown in Figure8.



Fig.8 Test results of uncivilized lying behavior

4. CONCLUSIONS

This paper presents a new data classification scheme for multi-object combination objectives and verifies and compares it by using multiple models. According to the experimental results, the careful analysis and processing of the data before the experiment can more accurately grasp the differences and similarities between all kinds of data, so that the algorithm model can have better detection performance.

However, the method of this paper also has some limitations. In the city, both public chairs and uncivilized poses are diverse. Limited by the lack of data, what the object studied in this paper objected at is the more typical public chair and uncivilized posture. For other types of uncivilized chair occupying behavior, its detection effect will be relatively poor. In addition, there is some room for optimization of the dataset classification scheme including annotation work.

ACKNOWLEDGMENTS

This work was supported by the High-Level Innovation Team Construction Project of Beijing Municipal Universities (No. IDHT20190506) and the BUCEA Post Graduate Innovation Project.

REFERENCES:

[1] TIAN D, ZHANG C, DUAN X, et al. The cooperative vehicle infrastructure system based on machine vision[C]//t The 6th ACM Symposium on Development and Analysis of Intelligent Vehicular Networks and Applications. New York: ACM Press, 2017: 85-89.

[2] FELZENSZWALB P F, GIRSHICK R B, MCALLESTER D, et al.Object detection with discriminatively trained part-based models[J].IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010,32(9): 1627-1645.

[3] MICHAL D,TOMASZ K. FPGA Implementation of multi-scale face detection using HOG features and SVM classifier[J]. Image Processing & Communications, 2016, 21(3): 27-44.

[4] Bao Xiaomin, Wang Siqi. Review of deep learning-based object detection algorithms [J]. Sensors and Microsystems, 2022,41(04):5-9.DOI:10.13873/J.1000-9787(2022)04-0005-05

[5] Wang, Jingyao and Naigong Yu. “UTD-Yolov5: A Real-time Underwater Objects Detection Method based on Attention Improved YOLOv5.” ArXiv abs/2207.00837 (2022): n. pag.

[6] Aveiro, João L. et al. “Identification of Binary Neutron Star Mergers in Gravitational-Wave Data Using YOLO One-Shot Object Detection.” ArXiv abs/2207.00591 (2022): n. pag.

[7] Keles, Muhammed Can et al. “Evaluation of YOLO Models with Sliced Inference for Small Object Detection.” ArXiv abs/2203.04799 (2022): n. pag.

[8] Huang, Dehao et al. “Bone Marrow Cell Recognition: Training Deep Object Detection with A New Loss Function.” 2021 IEEE International Conference on Imaging Systems and Techniques (IST) (2021): 1-6.

[9] ahman, Raian et al. “Densely-Populated Traffic Detection using YOLOv5 and Non-Maximum Suppression Ensembling.” ArXiv abs/2108.12118 (2021): n. pag.

[10] Zhu, Xingkui et al. “TPH-YOLOv5: Improved YOLOv5 Based on Transformer Prediction Head for Object Detection on Drone-captured Scenarios.” 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW) (2021): 2778-2788.

[11] Gu, Yongxiang et al. “Real-time Streaming Perception System for Autonomous Driving.” 2021 China Automation Congress (CAC) (2021): 5239-5244.

[12] Reddy, Rachala Rohith and Mahesh Raveendranatha Panicker. “Hand-Drawn Electrical Circuit Recognition using Object Detection and Node Recognition.” SN Comput. Sci. 3 (2022): 244.

[13] Ghiasi, Golnaz et al. “Simple Copy-Paste is a Strong Data Augmentation Method for Instance Segmentation.” 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021): 2917-2927.

[14] He, Kaiming et al. “Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition.” IEEE Transactions on Pattern Analysis and Machine Intelligence 37 (2015): 1904-1916.

[15] Lin T-Y，Dollár P，Girshick R，He K，Hariharan B，and Belongie S. Feature pyramid networks for object detection//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu，USA，2017：2117-2125

[16] Liu S，Qi L，Qin H，Shi J，and Jia J. Path aggregation network for instance segmentation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City，USA，2018：8759-8768

[17] Xie Xingxing, et al., " Dynamic feature fusion of remote sensing image object detection. Journal of Computer Science 45.04 (2022): 735-747.

[18] Li Shiwei. Object detection and license plate recognition system based on YOLOv5 algorithm. Electronic Technology and Software Engineering.01 (2022): 138-141.

[19] Pang J，Chen K，Shi J，Feng H，Ouyang W，and Lin D. Libra r-cnn：Towards balanced learning for object detection//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach，USA，2019：821-830

[20] Tan M，Pang R，and Le Q V. Efficientdet：Scalable and efficient object detection//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Seattle，USA，2020：10781-10790

[21] Liu S，Huang D，and Wang Y. Learning spatial fusion for single-shot object detection. arXiv preprint arXiv：1911.09516，2019

[22] Wang X，Zhang S，Yu Z，Feng L，and Zhang W. Scale-Equalizing Pyramid Convolution for Object Detection//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle，USA，2020：13359-13368

[23] BOCHKOVSKIY A, WANG C-Y, LIAO H-Y M. Yolov4: Optimal speed and accuracy of object detection[J]. arXiv preprint arXiv:2004.10934, 2020：2980-2988.