Behavior monitoring model of kitchen staff based on YOLOv5l and DeepSort techniques

Xiaotong Guo, *Min* Zuo*, *Wenjing* Yan, *Qingchuan* Zhang, *Sijun* Xie, and *Iker* Zhong

National Engineering Laboratory for Agri-Product Quality Traceability, Beijing Technology and Business University, Beijing, 100048, China

Abstract. Although the monitoring system has been widely used, the actual monitoring task still needs more manpower to complete. This paper takes yolov5l model and deep sort algorithm as the basic framework to identify and track the staff in kitchen environment. We apply a relation construction with detected items and people, then label the relation corresponding to behaviors violate the regulations of kitchen, such as the staff did not wear mask or hat. We train our model and the experimental results show that the model can correctly identify the inappropriate behaviors of staff. The model achieves the time-constrained accuracy of 95.32% in identifying whether the staff wear a hat or not, and the time-constrained accuracy of 96.32% in identifying whether the staff wear mask correctly. The result shows that the proposed model could fulfil monitoring task in this kitchen environment.

Keywords: Object detection, YOLOv5l model, DeepSort, Compressed deep learning model, Automation.

1 Introduction

Although the monitoring system has been widely existing, the actual monitoring task still needs manpower to complete. The existing video monitoring system usually only records video images, providing information without interpretation of video images, which can only be used for extracting evidence after the event. Recently the deep leaning algorithm has been used in target detection and recognition, which allows the compute to perform monitoring task automatically and intelligently. [1-8] It provides a certain basis for our research.

This paper focuses on the application of intelligent video monitoring in kitchen environment. In this paper, the algorithm based on yolov5l and deepsort is used to detect the people and items in the kitchen environment monitoring to identify whether the staff wear masks and hats correctly.

* Corresponding author: zuomin@btbu.edu.cn

2 Model

2.1 Yolov5l

The network structure of yolov5 is divided into four parts: input, backbone, neck and prediction. The input part completes the basic processing tasks such as data enhancement, adaptive image scaling and anchor frame calculation. In the backbone part, CSP (cross stage partial) structure is used to extract the main information from the input samples for subsequent use. The neck part adopts FPN (Feature Pyramid Networks) and PAN (Path Aggregation Network) structure, and uses the information extracted from the backbone part to enhance feature fusion.

In our model the lost function of prediction for bounding box applies GIOU_Loss

Is to make a prediction and calculate the loss value, such as GIOU_Loss. For two boxes A, B. Firstly, we calculate their minimum convex set (the minimum bounding box surrounding a and b) C. secondly, combined with the minimum convex set C, we calculate the formulas of GIOU and GIOU_LOSS as follows:

$$\mathrm{GIOU} = \mathrm{IOU} - \frac{\mathrm{C}-(\mathrm{A}\cup\mathrm{B})}{\mathrm{C}}, \tag{1}$$

$$\mathrm{GIOU\_LOSS} = 1 - \mathrm{GIOU}, \tag{2}$$

Yolov5l model updates yolov5 model in depth and width of the network construction. The backbone network part uses CSP structure three times with 3, 9 and 9 residual components. The neck part uses CSP structure five times, and yolov5l uses three residual components in each CSP structure.

2.2 Deepsort

Deep sort is a multi-target tracking algorithm. It uses motion and appearance information for data association. The algorithm detects the object in each frame, and matches the object with previous detection.

The weight of matched-degree is obtained by the weighted sum of Mahalanobis distance between the position and the similarity of the image with in the bounding box area. When calculating Mahalanobis distance, Kalman filter is used to predict the covariance matrix of motion distribution. The minimum cosine distance is calculated by using motion and appearance information. The matched-degree is defined in the following formula:

$$\mathrm{c_{i,j}} = \lambda \mathrm{d^{(1)}(i,j)} + (1-\lambda)\mathrm{d^{(2)}(i,j)}, \tag{3}$$

where $\mathrm{d^{(1)}}$ is Mahalanobis distance, $\mathrm{d^{(2)}}$ is cosine distance and λ is weight coefficient. The minimum cosine distance is calculated by using motion and appearance information.

2.3 Behavior regonition

Firstly, yolov5l is used to recognize objects (people, hats and masks), and then the bounding box of people are transferred into the system to detect violations.

This mode detects the behavior violation based on two rules:

Hat wearing: The hat wearing is identified based on the bounding box of people and hat. If the bounding box of hat locates above than the top quarter of the height of person's bounding box, it is defined as appropriate hat wearing.

Mask wearing: The inappropriate mask wearing, such as wear the mask on the chin is identified with “B mask” label. And not wearing mask is identified as label "C mask" .

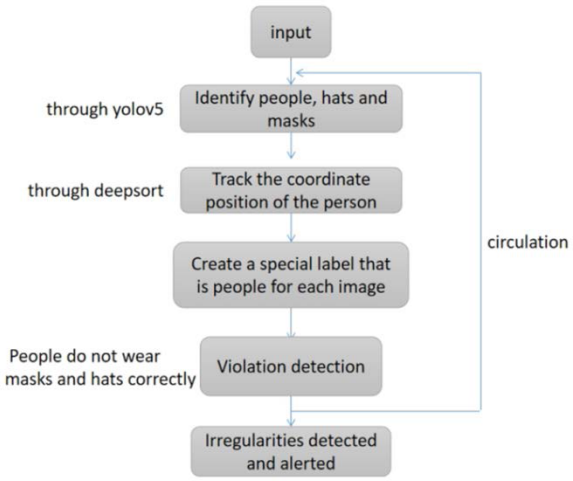

Fig. 1. Algorithm flow.

3 Experimental and results

3.1 Dataseting

The data source of behavior data set is collected in kitchen environment. A total of five different scenes is recorded. Each scene is recorded for an hour, and the total length of the video is 5 hours. The training dataset applies a total of 2000 pictures are captured in the recorded video. The item are labeled with five types, which are “person”, “hat”, “A mask”, “B mask” and “C mask”. Among them, label “person” refers to the location of the staff in the camera area. Label “hat” is the hat worn by the staff in the camera area. Label “A mask” is where the worker wears a mask in the camera area. Label “B mask” indicates that the staff does not wear the mask properly in the camera area. Label “C mask” means that the staff does not wear masks in the camera area. The labeling example is shown in the Figure 1.

Fig. 1. Annotation set picture.

3.2 Experimental implement

In this paper, in the process of the experiment, GPU is needed for calculation. Table 1 is the hardware environment configuration of this experiment.

Table 1. Description of the configuration of experimental hardware environment.

Product Name	Type
Processor	Intel Core i7-7700@3.60GHz
Memory	16GB (G.skill DDR4 2400MHZ)
Graphics card	NVIDIA GeForce RTX 3060
main board	Gigabyte b250m-evo-cf
monitor	AOC2479 2479w1 (23.8 in)
Main hard disk	CV3-8D128 (128GB/SSD)
network card	RTL8168/8111/8112 Gigabit Ethernet Controller
Sound card	Intel High Definition Auto Controller

3.3 Result

The performance of items recognition in kitchen environment is evaluated with labbelled dataset. This experiment takes 2000 labeled pictures, where 90% of them are used for training, and 10% of them are use for testing. The number of training iterations is set to 300, The results of the training data set are shown in Figure 2.

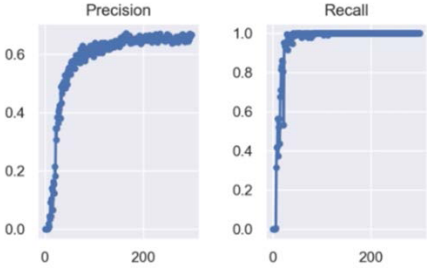

Fig. 2. The curve of precision and recall of training.

As can be seen from the Figure1, precision and recall gradually increase with the number of iterations. The final precision achieves 91.29%, while the final recall achieves 1.

The inappropriate behavior recognition is also evaluated. The model achieve accuracy of 95.32% in identifying whether the staff wear a hat or not, and the accuracy of 96.32% in identifying whether the staff wear mask appropriate. Accuracy is defined as the inappropriate behaviors in the video can be correctly recognized with the duration of half a minute. The example of inappropriate mask wearing is shown in Figure 3.

Table 2. Behavior recognition accuracy.

Construction	Behavior recognition accuracy/%
Hat	95.32
Mask	96.32

(a) (b)

Fig. 4. Violation detection. (a)original image (b)image after identification.

4 Conclusion

In this paper we proposed a hybrid model which combines yolov5l, deepsort and violation identification function. The model can effectively detect inappropriate behavior of the kitchen staff. Our research can can effectively reduce human labor in the task of the kitchen monitoring and realize the automatic supervision.

This study is supported by Beijing Natural Science Foundation (No.4202014), Natural Science Foundation of China (61873027), Humanity and Social Science Youth Foundation of Ministry of Education of China (No.20YJCZH229), the R&D Program of Beijing Municipal Education Commission (No.KM202010011011).

References

1. He K, Gkioxari G, Dollár P, et al. Mask r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2961-2969.
2. Koirala A, Walsh K B, Wang Z, et al. Deep learning for real-time fruit detection and orchard fruit load estimation: Benchmarking of ‘MangoYOLO’[J]. Precision Agriculture, 2019, 20(6): 1107-1135.
3. Peña-Barragán J M, Ngugi M K, Plant R E, et al. Object-based crop identification using multiple vegetation indices, textural features and crop phenology[J]. Remote Sensing of Environment, 2011, 115(6): 1301-1316.

4. Yan B, Fan P, Lei X, Liu Z, Yang F. A Real-Time Apple Targets Detection Method for Picking Robot Based on Improved YOLOv5. Remote Sensing. 2021; 13(9):1619.
5. Li H, Deng L, Yang C, et al. Enhanced YOLOv3 Tiny Network for Real-Time Ship Detection From Visual Image[J]. IEEE Access, 2021, 9: 16692-16706.
6. Girshick R. Fast r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2015: 1440-1448.
7. Laroca R, Zanlorensi L A, Gonçalves G R, et al. An efficient and layout-independent automatic license plate recognition system based on the YOLO detector[J]. IET Intelligent Transport Systems, 2021, 15(4): 483-503.
8. Tian Y, Yang G, Wang Z, et al. Apple detection during different growth stages in orchards using the improved YOLO-V3 model[J]. Computers and electronics in agriculture, 2019, 157: 417-426.
9. Wu W, Yin Y, Wang X, et al. Face detection with different scales based on faster R-CNN[J]. IEEE transactions on cybernetics, 2018, 49(11): 4017-4028.