10.18686/aitr.v2i2.4014

Multi-Scenario Mask Detection Method and Implementation Based on Deep Learning

Mingwei Han

North China University of Technology, Beijing, 100144

Abstract: Object detection is one of the core issues in the field of computer vision, with the task of accurately locating and recognizing all objects of interest in images, determining the categories and positions of objects. Regarding masks as targets in object detection, utilizing deep learning techniques to detect the wearing of masks on faces can greatly improve issues such as high manual supervision costs, low efficiency, and subjective differences. In this paper, based on the YOLOV5 algorithm, we propose a mask detection technology that can accurately detect the wearing of masks in real-time, and experimentally determine the occlusion boundaries.

Keywords: Deep learning; Target detection; Real-time detection of the YOLOv5 algorithm

Introduction

After the end of the COVID-19 pandemic in the first winter, with the fluctuating temperatures and continuous virus mutations, children's health faces severe challenges: ongoing infections such as: Mycoplasma pneumoniae infection, respiratory tract infections, and others. In the face of infectious diseases, the most effective method is wearing masks to reduce gatherings. The rapid development of deep learning in daily life has increased the level of intelligence, achieving breakthrough progress in various applications such as computer vision, natural language processing, speech recognition, recommendation systems, and medical fields.[1]

1. Current Research Status at Home and Abroad

1.1 Research Status of Object Detection Based on Deep Learning

The application of convolutional neural networks (CNNs) in deep learning-based object detection algorithms involves the automatic extraction of target features. The development process can be categorized into Two-Stage and One-Stage algorithms based on the different detection processes.

The Two-Stage algorithm initially extracts candidate regions in the image that may contain targets, followed by regression within these candidate regions to obtain detection results. In 2014, Girshick et al.[2] first proposed the R-CNN object detection network, namely the Region-based Convolutional Neural Network (R-CNN), which marked the first appearance of deep learning-based algorithms in the field of object detection. In 2015, Girshick et al.[3] introduced the Fast R-CNN object detection network, which optimized the grid structure and integrated the advantages of R-CNN and SPPNet. This network achieved synchronous target classification and candidate box regression, reducing the training time to 9.5 hours and achieving a mean Average Precision (mAP) of 66.9%.

One-Stage algorithms directly obtain specific position and category information of the target to be detected through regression.

In 2016, Joseph et al. proposed the YOLO (You Only Look Once) algorithm, which completely abandoned the "region proposal + regression" mode of two-stage detection and instead adopted a one-stage end-to-end processing approach. Its most significant advantage is its fast detection speed. However, this model may exhibit slightly lower detection accuracy. At the end of 2016, Redmon et al.[4] troduced the YOLOv2 algorithm, which utilized Darknet-19 as the backbone network and introduced batch normalization layers in each convolutional layer to achieve regularization effects. In 2018, Redmon et al.[5] proposed YOLOv3, which introduced residual network modules based on Darknet19 and utilized a more powerful Darknet53 as the backbone network for image feature extraction. In 2020, Bochkovskiy et al.[6] introduced the YOLOv4 algorithm, which incorporated Mosaic data augmentation to enrich the dataset and reduce computational costs simultaneously. The YOLOv5s model is compact, fast, and accurate, suitable for real-time detection and deployable on mobile devices.

1.2 Current Research Status of Mask-wearing Detection

Many researchers have been continuously refining mask-wearing detection algorithms based on deep learning techniques, leading to the emergence of a plethora of efficient and accurate solutions.

Li Yuyang et al. [7] merged the classification scores and IoU scores of the Feature Pyramid Network (FPN) and the Conjugate Gradient Descent (CGD) algorithm based on the SSD algorithm. They introduced QFL loss into the loss function, resulting in a decrease in detection speed by 2.4 frames/s compared to the original SSD algorithm. Wan Zilun et al. [8] improved the Faster R-CNN algorithm in four aspects: selecting ResNet-50 convolutional neural network as the image feature extractor, enhancing the multi-task enhanced RPN model. The mean precision reached 90.18%, but issues such as video lag and distortion occurred during real-time video detection, indicating insufficient model efficiency.

This study leverages the YOLOv5 algorithm to achieve real-time monitoring of mask-wearing conformity. The aim is to accurately and swiftly identify whether mask-wearing is compliant in various complex scenarios. The main contents are outlined as follows:

Chapter 1: Introduction. This chapter presents the research background and significance of mask-wearing detection algorithms, reviews the research status of object detection and mask detection domestically and abroad.

Chapter 2: Data Acquisition and Annotation. This chapter provides detailed information on the self-made dataset used in this study and its preprocessing methods.

Chapter 3: Analysis of Test Results. This chapter primarily analyzes the detection results of the model in three different scenarios. The performance of the model is evaluated, and a systematic observation of the change in facial occlusion area from small to large is conducted.

2. Experimental Design and Analysis

2.1 Introduction to the Dataset

Given the relative scarcity of samples for mask-wearing detection in current publicly available datasets, along with issues such as mislabeling or omissions in some images, this study constructed a mask detection dataset. The dataset comprises a total of 7267 images, primarily selected from the open-source MAFA dataset, supplemented by images obtained through web searches. Specifically, we extracted 6000 images depicting mask occlusion from the MAFA dataset.

In this study, a total of 1267 additional mask images were collected from the internet to supplement the dataset, including images of children wearing masks and instances of masks with multiple occlusions.

2.2 Dataset Processing

2.2.1 Image Labelling

For the unlabeled and self-collected data in the dataset, the LabelImg tool was utilized to manually annotate these images. Each label corresponds to a numerical index, where index 0 represents "masked", index 1 represents "unmasked", and index 2 represents "nonstandard-masked". Nonstandard-masked indicates improper mask-wearing, such as exposing the mouth, nose, or chin while wearing a mask. The labeling process involved opening the annotation interface using relevant commands in the terminal, saving the annotations in the specified folder according to the VOC annotation file path, and storing the annotation information as XML files.

2.2.2 Format Conversion

MAFA data sets are formatted for category annotation, and the corresponding annotation information is stored in the form of XML files. However, in the YOLO algorithm, in order to facilitate the input and processing during the training of the YOLO algorithm and improve the call efficiency during the training, the label file in xml format needs to be uniformly converted into the TXT format of the simpler YOLO format through code. The converted TXT file format is shown in Figure 2.1.

```
1 0.1075 0.27982954545454547 0.11 0.11647727272727273
0 0.4025 0.2911931818181818 0.135 0.14488636363636365
0 0.78125 0.19318181818181818 0.1125 0.14204545454545456
```

Fig. 2.1 TXT file annotation format

A total of 2094 images were manually labeled in this study. After the data set annotation was completed, there were a total of 7267 images, among which 3516 targets were wearing masks and 2136 targets were not wearing masks. The target number of irregular wearing masks was 1615. At the same time, in order to improve the training effect of the model, the whole data set is divided into training set, verification set and test set by 8:1:1.

2.3 Evaluation Model Index

Accuracy (A), accuracy (P), recall (R), average accuracy (AP) and detection frame rate per second (FPS) are generally used in the field of target detection. In this paper, the Yolo V5 algorithm was used to train 7267 pictures for 150 rounds, and F1 curves were drawn again to evaluate the performance of the trained model. F1 score was F-Measure, which integrated the two indicators of model precision and recall. To help determine how the model performs at different thresholds. It provides a measure of the model's balance between positive and negative cases, taking into account the accuracy and comprehensiveness of the model, and its formula is as follows:

$$F1 = \frac{2*P*R}{P+R}$$

According to the F1 curve shown in Figure 2.2, it can be observed that the F1 values of labels "masked", "nonstandard-masked", "unmasked" and "old man" are more significant, and the overall trend of the curve is more stable, indicating that the model has a better recognition effect on these labels.

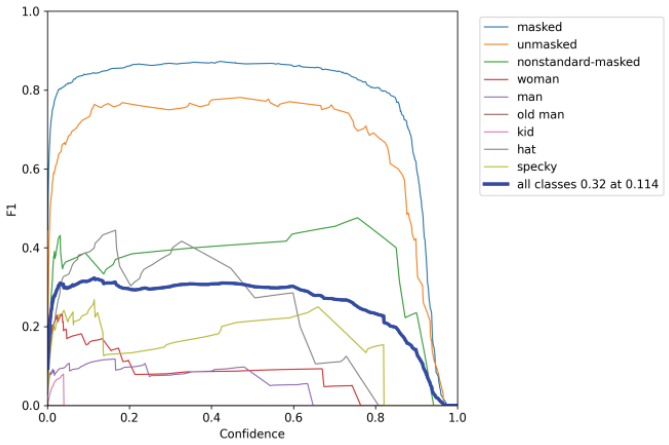


Fig. 2.2 F1 curve for each detection type

3. Experimental results and analysis of mask detection

3.1 No obvious occlusion, mask detection under good lighting conditions

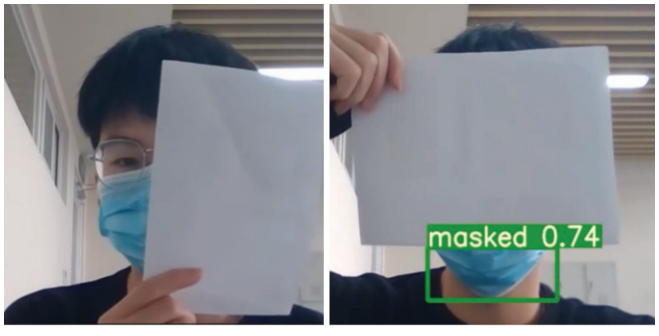
Under normal circumstances, the model can clearly identify whether the target is wearing a mask properly and give the corresponding probability. The recognition results are shown in Figure 3.1.

Fig. 3.1 Detection result in common scenarios

It can be obviously observed from the figure that the YOLOv5 model shows excellent accuracy and reliability in the identification of mask wearing conditions, and the model also shows excellent detection ability in multi-person scenarios, which can effectively deal with various complex situations.

3.2 Mask detection in the case of partially occluded face

In this section, we focus on the occlusion boundaries of the model. Turn on the live camera and use white paper to block different areas and from different angles. The results of our study are shown below:



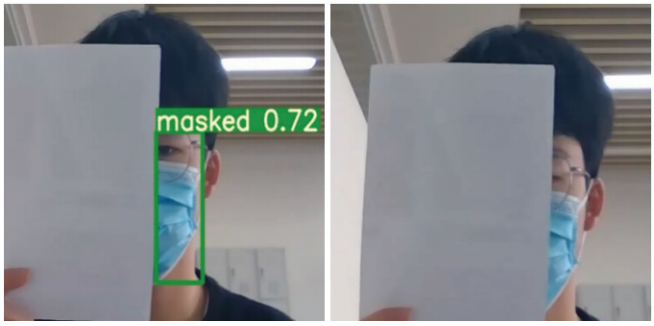


Fig. 3.2 Facial occlusion test results

The results show that the occlusion boundary is roughly four-fifths of the head area, meaning that when the occlusion exceeds this limit, the model will not be able to accurately determine whether to wear a mask.

References

[1] Sun Zhijun, Xue Lei, Xu Yangming, et al. Application Research of Computers, 2012, 29(8): 2806-2810. (in Chinese)

[2] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 580-587.

[3] Girshick R. Fast r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2015: 1440-1448.

[4] Redmon J, Farhadi A. YOLO9000: better, faster, stronger[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 7263-7271.

[5] Redmon J, Farhadi A. Yolov3: An incremental improvement[J]. arXiv preprint arXiv:1804.02767, 2018.

[6] Bochkovskiy A, Wang C Y, Liao H Y M. Yolov4: Optimal speed and accuracy of object detection[J]. arXiv preprint arXiv:2004.10934, 2020.

[7] LI Yuyang, SHEN Jiquan, ZHAI Haixia, FENG Weihua. Mask Wearing Detection Algorithm Based on Improved SSD[J]. Computer Engineering, 2022, 48(8): 173-179, 186.

[8] WAN Zilun, ZHANG Yanbo, WANG Duofeng, SUN Yichen, GU Fengyang, CHEN Mingyue. Face mask detection algorithm for multi task recognition in complex environment[J]. Microelectronics & Computer, 2021, 38(10): 21-27. DOI: 10.19304/J.ISSN1000-7180.2021.0056