Human Detection Based on YOLOv5 Network

Wenbo Zhong *

CATS College Canterbury, Canterbury, UK

* Corresponding Author Email: wz118235@catscanterbury.com

Abstract. With the vigorous development of the economy, the improvement of people's quality of life, the continuous increase in the ownership rate of family cars, and the increasing pressure on traffic management, people urgently hope to use new technologies to improve and optimize the past system. Therefore, the research on the modernization and intelligence of traffic control has become a research hotspot for scholars. When the crowd density is too high, all kinds of accidents are easy to happen, and it will bring difficulties to the management of order managers. Timely detection of crowd density, and when the crowd density is too high, warning information will be issued, which is conducive to the managers to evacuate the crowd in time. So, the final project is about how to recognize humans in the image, and with men and women. Today's human detection can be broadly divided into traditional image processing methods and machine learning-based methods. The traditional image processing method detects the human by preprocessing the image and then detecting the human contour. Frequent stampede events in large events at home and abroad have caused a lot of casualties. Therefore, the research on crowd counting has attracted more and more attention. If the crowd density of the current scene can be accurately estimated and the corresponding security measures can be arranged, the occurrence of such incidents can be effectively reduced or avoided. Crowd counting is widely used in video surveillance, traffic monitoring, public safety, urban planning, and the construction of intelligent shopping malls.

Keywords: Neural Network, Human Detection, Deep Learning.

1. Introduction

Rapid detection of crowd density, and when crowd density is too high, warning information will be provided, allowing managers to evacuate the crowd in time. Current human detection can be divided into two broad categories: traditional methods of image processing and methods based on machine learning. The traditional method of image processing detects people through certain image preprocesses and then detects the human outline. This method usually has low detection accuracy, and there will be a lot of missed detection. Nowadays, some methods based on machine learning have a good effect in detecting scattered crowds, but it is difficult to detect dense crowds. There is no doubt that self-driving travel has become the main way of travel for the people. The popularity of cars not only facilitates people's life but also brings many new jobs and increases the direction of people's employment. And with the rising of Automatic driving vehicles disappearing, object detection become quite significant. It is also reported that the penetration of electrical vehicles will increase from 4.6% (IEA) in 2020 to more than 50% in 2030, in which autonomous driving technologies are widely used [1]. As one of the key technologies in autonomous driving, video vehicle detection to detect vehicles driving on the roads and recognize their models, colors, and orientation, etc., has received increasing attention from both academia and industry.

When the crowd density is too high, all kinds of accidents are easy to happen, and it will bring difficulties to the management of order managers. Timely detection of crowd density, and when the crowd density is too high, warning information will be issued, which is conducive to the managers to evacuate the crowd in time [2]. The label is one of the key data in supervised learning. When coming to a new field, the first thing that comes to the mind is often what the label is, the process of labeling the dataset, and what the model should learn. The label in the crowd count is the density map, and the labeling process is mainly divided into two parts: (1) Crowd image annotation representation; (2) The crowd image annotation is converted into a crowd density map.

The main task of object detection is to detect specific objects in computer images. At the same time, it can also calculate the order in the given image according to the need target size and location, target detection in security, military, medical, transportation, life, etc. It is of great research value in the field of science

and technology. After searching for information found out that this study can use the website called MakeSense to create labels and takes this into training [3]. That's creating a custom model to detect the objects is an iterative process of collecting and organizing images, labeling the objects of interest, training a model, deploying it into the wild to make predictions, and then using that deployed model to collect examples of edge cases to repeat and improve.

The rest of this paper is organized as follows: Section 2 mainly discusses with dataset and YOLO algorithm, more details, and data. And Section 3 provides results and discussion, and Section 4 is about the conclusion.

2. Methodology

2.1. Data description

First, this study needs to collect some information and dataset for the basics. This study needs about 1,000 images spread in two different categories, male and female. The test has 200 images, and the training has 800 images. Figure 1 shows the sample images.



Figure 1. Illustration of the images of males and females.

The first step is to create labels, this paper prepared about a thousand images and use about 800 images to training and take labels. Figure 2 shows the tool that need to deal with [3], and specific step of how to create labels in Figure 3-5.

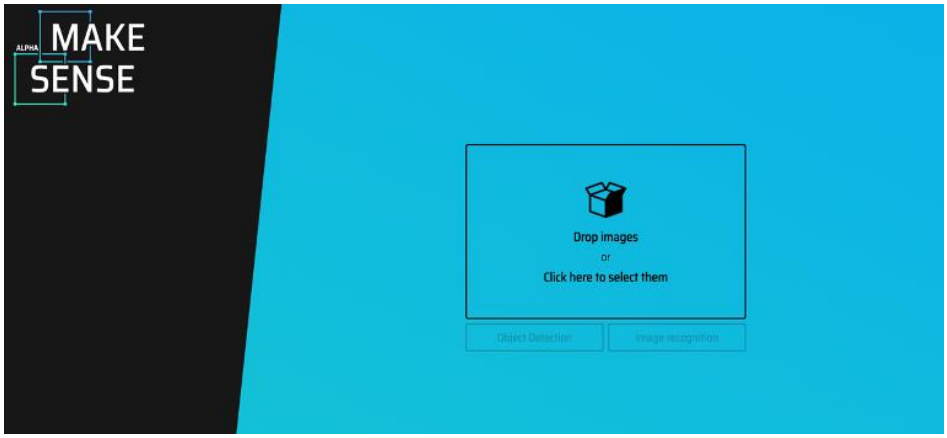


Figure 2. Illustration of tool 'MakeSense'.

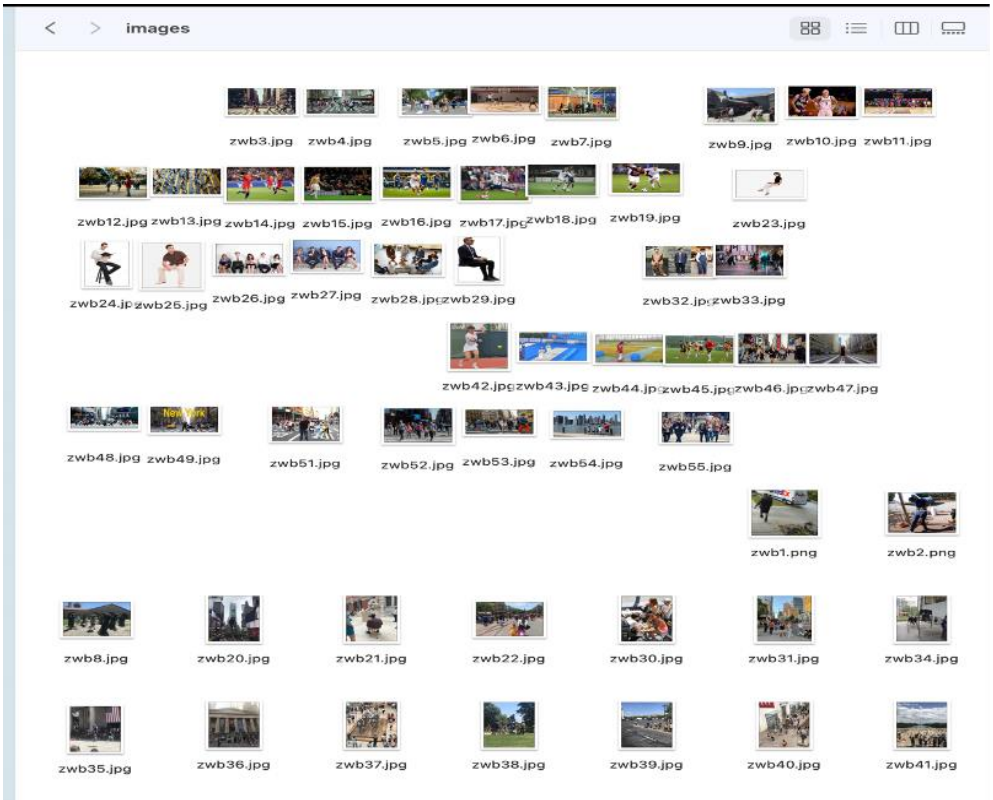


Figure 3. Illustration of images example.



Figure 4. Illustration of steps of create labels.

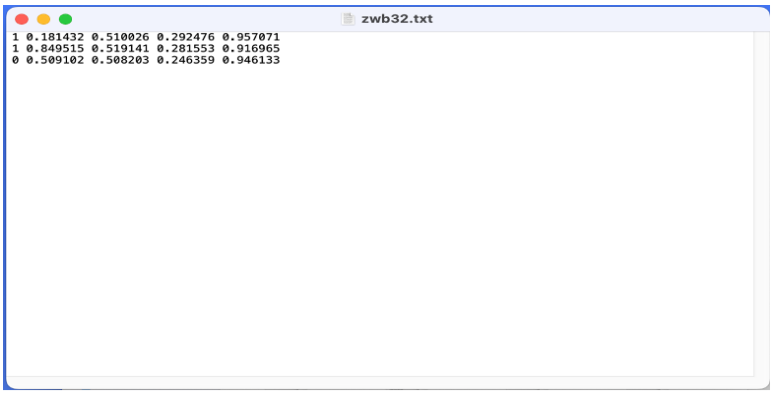


Figure 5. Illustration of the YOLOv5 labels in .TXT format.

The integer numbers in the first column of the .TXT file refer to the object class id, and. The class id 0 and 1 represent the “men” and “women” classes in this dataset. In each row, the second and the

third decimal numbers indicate the x and y coordinates of the center of the bounding box, and the last two numbers denote the wide and the height of the bounding box. As it shows, these coordinates are normalized to [0, 1].

YOLOv5 reads a .TXT file. This requires us to convert the XML file into a .TXT file. Through the online annotation tool Make Sense, you can directly save the image annotation file in .TXT format. Just need to circle the person in the picture.

2.2. Neural Network

The neural network is an important branch of machine learning and an important basis for deep learning. As one of the most important artificial intelligence technologies, it not only drives the hot development momentum of machine learning but also promotes the rapid development of deep learning technology, which is widely used in many tasks [4-6]. In the early 20th century, neurobiological studies found that the cerebral cortex can perform specific conversion of input brain signals and then output signals to other cortexes, as shown in Figure 6. The cerebral cortex is composed of a large number of neurons and glial cells supporting neurons, which is closely related to human intellectual function. Neurons, as the basic information processing unit, have become the focus of research. In 1943, McCulloch et al. proposed a mathematical logic model of neurons. Therefore, the study of the artificial neural network is more and more in-depth. Inspired by the function of biological neural networks, researchers have the idea of using neural networks to construct machine systems to simulate human intelligent behavior, which is the bud of the artificial neural network method. With that, the rapid development of the neural network, the theory gradually tends to mature, a variety of new networks like bamboo shoots after a spring rain, and the neural network family is growing. Traditional machine learning mainly focuses on artificially designed feature extraction algorithms, which are highly subjective, while the deep learning models using neural networks have strong self-learning feature extraction ability, high universality, and high ductility.

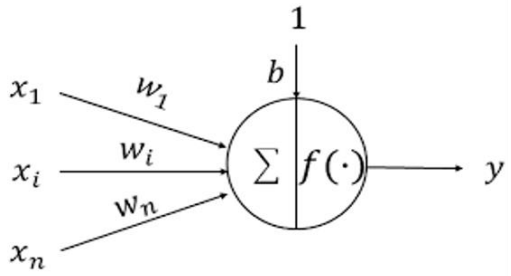


Figure 6. Illustration of the architecture of the neuron model [7].

2.3. CNN

At the end of the 20th century, LeNet-5 network came into being. However, due to the low processing power of computers at that time and the lack of data sets for training, the ability of convolutional neural network to deal with complex problems has not been shown, so scholars have not paid attention to it. It was not until more than ten years later, when Hinton and his students won the first prize in the image classification competition with AlexNet network shown in Figure 7, that convolutional neural network came back into people's view. The rapid development of computer processing power and the continuous progress of chip technology have re-stimulated the research upsurge of convolutional neural networks, and deep learning has entered a period of rapid development both academically and industrially. The unique advantages of convolutional neural network make it one of the most important and famous networks at that time. For example, in the face of large images, convolutional neural networks are more likely to grasp the spatial information features, which alleviates the problem of easy loss of spatial information in fully connected neural networks to a certain extent. In addition, the convolutional neural network reduces a large number of useless parameters, adopts the parameter sharing strategy, the training process is faster, and has a greater application prospect.

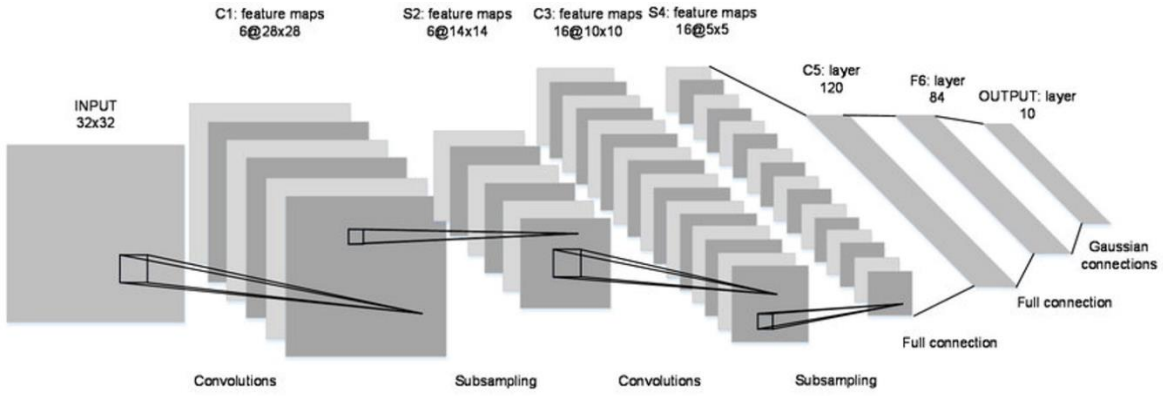


Figure 7. Illustrations of architecture of the convolutional neural network (CNN) model.

The basic structure of convolutional neural network includes input layer, convolutional layer, pooling layer, fully connected layer and output layer [8]. The input layer is used to accept input data; As the core layer, convolutional layer is mostly used for feature extraction and undertakes the most computational engineering. The more network layers there are, the more advanced the extracted features will be. Pooling layer is used for pixel sampling, also known as sampling layer, which can reduce the complexity of the network. Common operations include MaxPool (maximum pooling), AveragePool (mean pooling), etc. Through continuous convolution pooling operation, image features can be abstracted into high-dimensional space. The fully connected layer is usually used for the final integration of feature information after multiple convolutional pooling operations, and passed to the output layer through the activation function. The output layer is responsible for the final output, which varies according to the specific application, and can be classified categories, predicted values, and so on. In computer vision, traditional methods are more dependent on the quality of feature extraction methods, with great subjectivity, and it is difficult to excavate some deep features. However, convolutional neural network adopts adaptive learning to train the ability to extract useful features, which can improve the efficiency of recognition to a certain extent.

At present, the common convolutional neural networks include AlexNet, VGG, ResNet [9], etc.

2.4. Introduction for YOLO

YOLO is the first work of One-stage Detection, with the full name of "You Only Look Once". As the name implies, it can complete its work within a glance. The conventional object detection algorithm is divided into two steps: candidate and detection. Although it has a good level of accuracy, its speed is always unsatisfactory. YOLO creatively combines the two steps into one, allowing you to accurately identify the location and type of object at a glance. The design concept of YOLO network comes from GoogleNet. The whole network is composed of 24 convolution layers and 2 fully connected layers. The difference from GoogleNet lies in the common use of 3x3 and 1X1 convolution kernels. YOLO combines recognition and location into one. After the convolutional layer extracts the image features, the fully connected layer directly predicts the image position and category probability value, with simple structure and fast detection speed. Figure 8 shows the different scales of network parameters of YOLOv5.

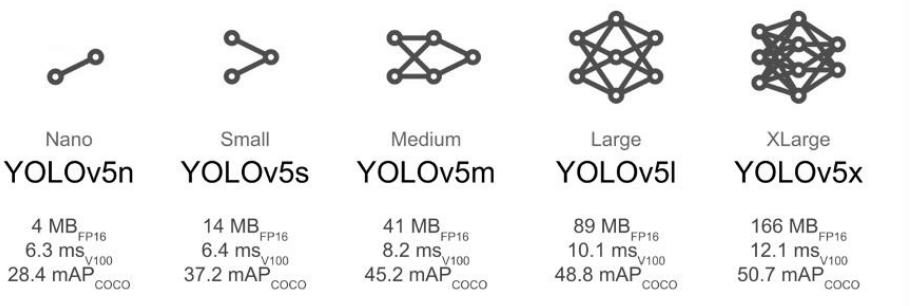


Figure 8. Different scales of network parameters of YOLOv5 [10].

Yolov5 Network structure diagram:

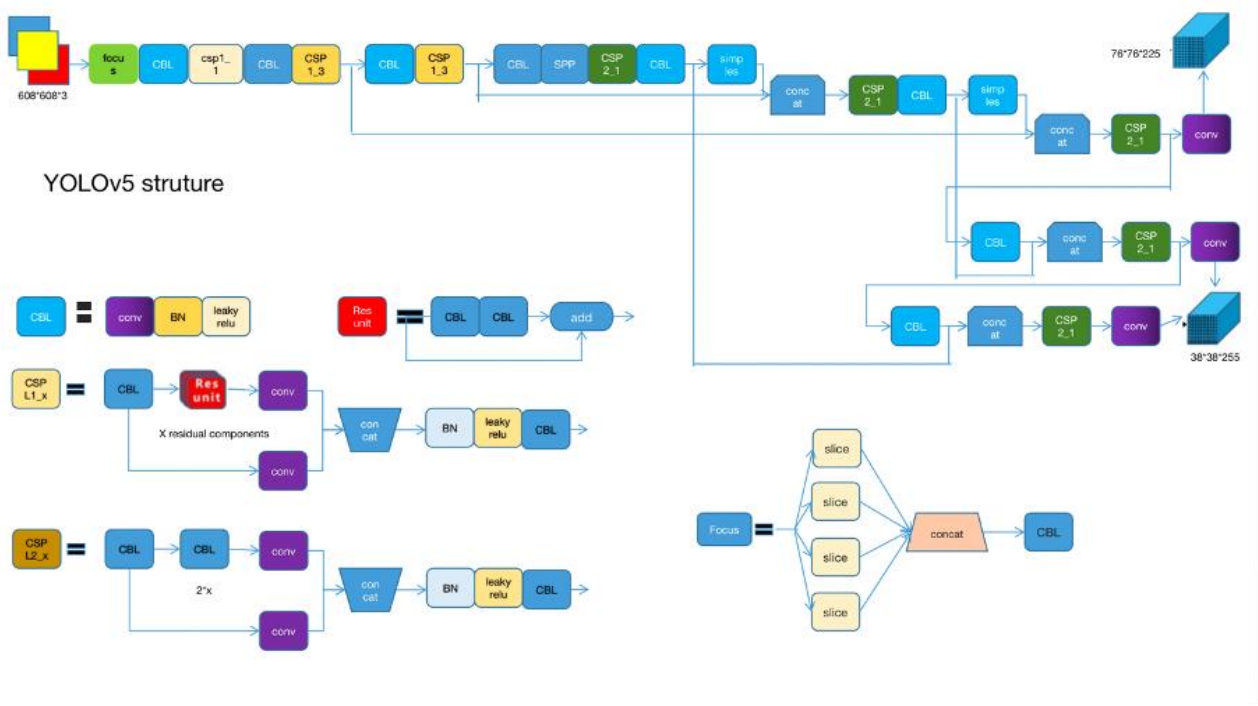


Figure 9. Illustration of the architecture of the YOLOv5.

Figure 9 illustrates the network architecture of Yolov5. It consists of three parts: (1) Backbone: CSPDarknet, (2) Neck: PANet, and (3) Head: Yolo Layer. The data are first input to CSPDarknet for feature extraction, and then fed to PANet for feature fusion. Finally, Yolo Layer outputs detection results (class, score, location, size).

Backbone is a convolutional neural netowrk that extracts image features in different granularities and concatenates them together. Neck is a stack of network layers that mix and combine image features and pass them to the prediction layer (usually FPN or PANET). Head is use to predict image features, generate bounding boxes and predict categories.

3. Results and Discussion



Figure 10. Illustration of human and gender detection.

3.1. Results

Here is an example that this paper are going to take in this case, this project labeled all clearly visible people in the picture shown in Figure 10, to prevent missed judgements from the system later on, and decided to only keep a record of all human that has been blocked minorly. And in this picture

everyone has been labeled by a box, and it is either male or female. Kept doing the same process for about 800 pictures.

Human detection: As one of the most critical steps, human recognition requires strong detection capability of the network, which can reduce false detection rate. In this study, YOLOv5 model with high speed and precision is used for recognition, and the pre-trained model on COCO dataset is still used for transfer training.

Figure 11 shows some obtained evaluation results including the changes of Loss, box_loss, object_loss, cls_loss in the training process and precision and recall.

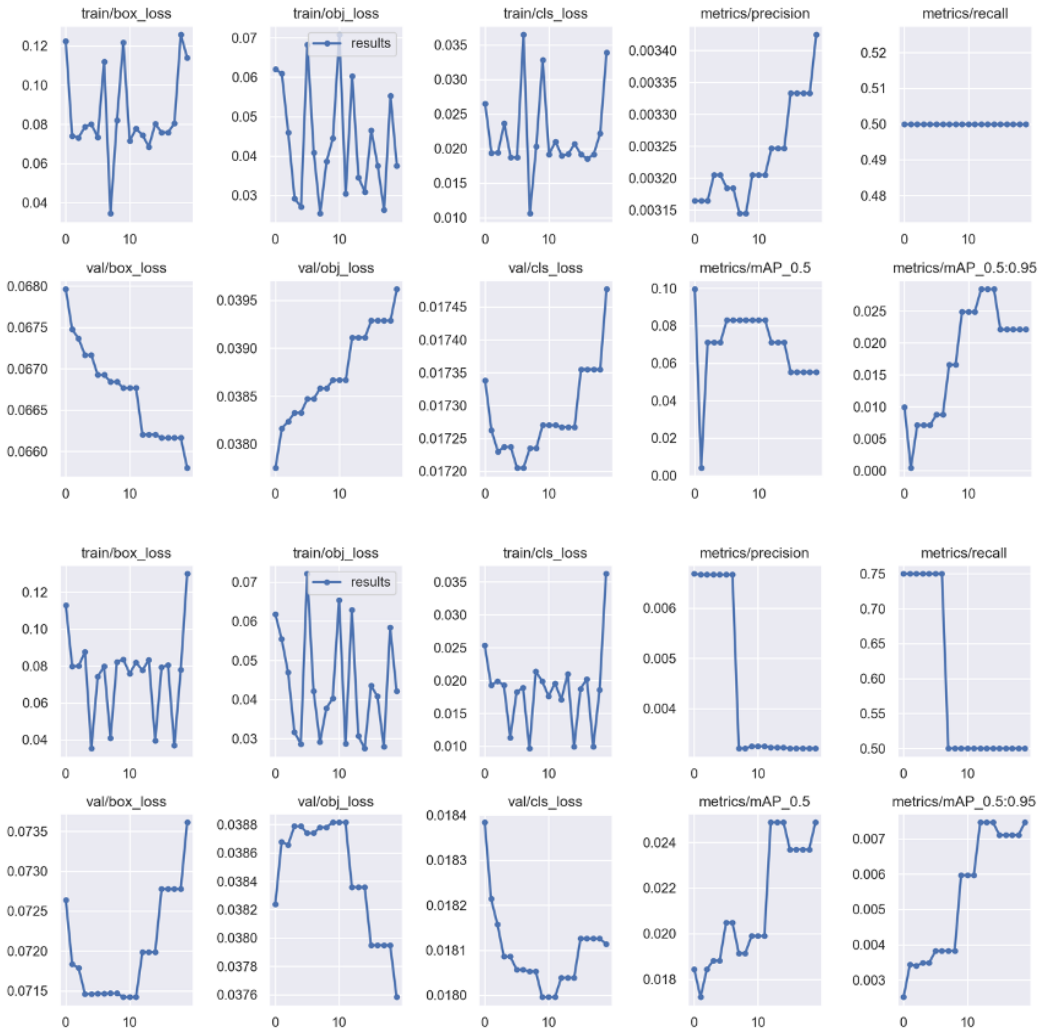


Figure 11. Illustration of change of loss and other evaluation metrics.

Precision surge as the increase of the training iterations and metrics recall kept stable, and val/box's loss decreases as the increases of training, and object's loss increased at first and dropped in the first image.

Precision and recall declines rapidly in the first a few round then levels flatten out and finally converges at 0.001 and 0.50 in the image (2).

3.2. Discussion

In total, program have a high percentage of successful tests up to 92.5%, with 5.5% of failure and 2% of misjudgements.

By reviewing the results of all fracture tests, this study came up with three main reasons for misidentification. The first is objects with human-like features, such as the same structure, the same color, the same clothing. The second is when parts are missing, when a person's arm, leg or head gets stuck, and the system fails to recognize them as a person. The third is gender error. According to the

test results, most errors occur when men have long hair and women have short hair, which leads the system to make wrong judgments.

That, however, is another test result. The scarecrows were identified as human. This happens because scarecrows have similar attributes to humans, and scarecrows are also used to pretend to be human. So the system didn't really make a mistake, but in theory, it made an error of judgment, shown in Figure 12.

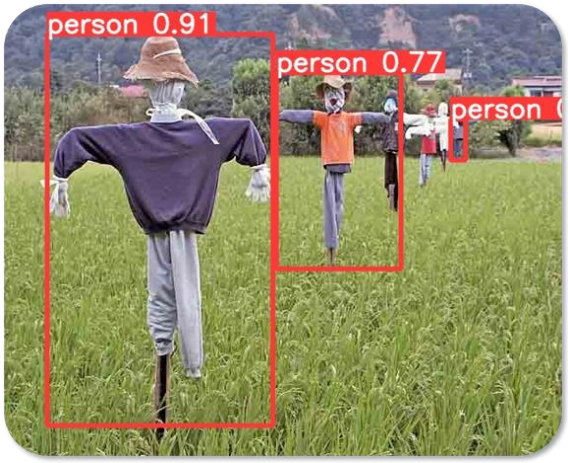


Figure 12. Illustration of different test.

4. Conclusions

In this paper, to comprehensive analysis and understanding of human detection and gender detection and today's all kinds of deep learning networks in the application of human detection. Just like crowd counting is a relatively small CV direction. At present, human detection uses deep learning and recognizes people. Based on the analysis of the experimental results, this study has the following shortcomings: 1. In the process of training graphics, a better GPU is needed to make the work more efficient. 2. The content of identification is limited, and the data sample is small. Further tests can be added in the future. Experimental results showed that elaborated the real-time Total number of errors identified tests up to 92.5%, with 5.5% of failure and 2% of misjudgments. In summary, image detection has made huge progress, although there is mature image recognition in image detection, there is still a long way to go and too mature results have not been obtained in detection. It still has a very big development prospect in the future.

References

[1] Zhao J. et al. Research and implementation of license plate recognition technology [C]. 2012 24th Chinese Control and Decision Conference (CCDC). IEEE, 2012.

[2] Guo Z. A real-time monitoring method and process of crowd density based on YOLOv5 [P]. 2021 http://www.xjishu.com/zhuanli/55/202011551457.html

[3] Makesense.ai [R], 2022. https://www.makesense.ai/

[4] Kayalibay B et al. CNN-based segmentation of medical imaging data [R]. arXiv preprint arXiv:1701.03056 (2017).

[5] Qiu Y, et al. Pose-guided matching based on deep learning for assessing quality of action on rehabilitation training [J]. Biomedical Signal Processing and Control 72 (2022): 103323.

[6] Qiu T et al. Efficient medical image enhancement based on CNN-FBB model [J]. IET Image Processing 13.10 (2019): 1736-1744.

[7] Datatron. A beginner friendly explanation of how neural networks work [J]. 2022, https://datatron.com/a-beginner-friendly-explanation-of-how-neural-networks-work/

[8] Wiki. Layers of a convolutional neural network [J]. 2022. https://wiki.tum.de/display/lfdv/Layers+of+a+Convolutional+Neural+Network

[9] Ou J et al. Vector-kernel convolutional neural networks [J]. Neurocomputing, 2019, 330: 253-258.

[10] Github. Train Custom Data [R]. 2022. https://github.com/ultralytics/yolov5/wiki/Train-Custom-Data