Object detection and recognition system based on computer vision analysis

Haitao Liu1*, Yuge Li 2 and Dongchang Liu1

1 Institute of Automation, Chinese Academy of Sciences, Beijing, China
2 Renmin University of China, Beijing, China

*Corresponding author. Email: liuhaitao2018@ia.ac.cn

Abstract. Artificial intelligence based on deep learning enables the machine to have the ability of understanding and cognition, but the application of artificial intelligence technology in supermarket shopping scene is limited. In the post-epidemic era, the contactless self-checkout of unmanned supermarket is more in line with the development needs of modern society. We build Pytorch environment, first to collect pictures of a large number of commodities and labeling information, and training model is obtained by YOLO neural network algorithm, finally through a call to model to realize the recognition of goods. Neural network algorithm is used to improve the recognition rate of goods step by step and achieve the detection and recognition of objects. We have tested our model on the real supermarket commodity data set and the public data set ImageNet, and the results show that our model can achieve a certain practical effect.

Keywords: post-epidemic, artificial intelligence, computer vision, deep learning.

1. Introduction
Artificial intelligence has become a hot topic in the world. In recent years, life is also changing drastically with the development of AI technology. Domestic artificial intelligence has also been developing. The State Council has issued a series of artificial intelligence development plans, which can be seen that the country attaches great importance to the development of artificial intelligence. Due to the impact of the global epidemic, the demand for contactless shopping is more urgent, and unmanned supermarkets are developing rapidly.

Computer vision in deep learning networks is the most rapidly developing branch of artificial intelligence. Now that we've entered an age of pictures and videos, the information explosion is due to the Internet as a carrier of information, and in part because of the sheer number of sensors of all types. Visual information is the most difficult information to use, and we can infer the relationship between these information through mathematical models. On the Internet, it's hard to get specific information about them. To take a very simple example, right now YouTube's servers receive over 150 hours of video uploads every 60 seconds, and that's just 60 seconds. There's so much data that we can't look at it in its entirety and label and classify it with our eyes. Now the YouTube team or Google are trying to figure out how to tag, categorize, index and so on this data for advertising purposes or to help us retrieve or manipulate this data. It didn't work, because no one could handle such a huge amount of data by hand. The only hope for the job was computer vision, which could tag and categorize photos,
process frames in video, and automatically capture a great basketball shot. Computer vision is a subject closely related to many fields, and the visual models now established are also based on many fields [9], such as biology, psychology, engineering and mathematics.

Since China entered the 21st century, the Internet economy has developed rapidly, and the construction of high-speed network infrastructure has also been carried out in an all-round way, which provides the premise and guarantee for the realization of various online functions. With the rapid development of the Internet, China has also achieved rapid development in the construction of a cashless society and economy. Many people have not even applied for credit cards, but have directly transitioned to the stage of virtual currency transactions, which benefits from the common recognition and efforts of the whole society [1]. On this basis, a series of potential social needs have been developed, unmanned supermarket has also been rapid development. Unmanned supermarket covers an area of small, and the shopping process of self-help, from the waiter to settle accounts and check, to achieve the real unmanned. This kind of shopping mode will definitely develop and mature in the next few years. In the view of the whole society, people have great expectations for its development [10].

![Diagram of system structure](image_url)

**Figure 1. System structure.**

In this paper, we realized the automatic detection of commodities through deep learning neural network research, and completed the reduction of human cost, which can better meet the needs of social development in the post-epidemic era.

2. Related work

With the rapid development of information science, computer vision in the field of artificial intelligence also shows strong vitality. (1) At present, target vision based on learning and spatial vision based on 3D continue to be "mutually independent". Deep learning can hardly replace geometric vision in a short time. (2) The target location based on computer vision will gradually trend to "application research", especially reflected in the multi-sensor computer vision location fusion technology [6]. (3) As for the target object recognition, the target object recognition based on deep learning will develop from "general recognition" to "object recognition in specific domain". "Domain Specific" can provide more accurate and specific prior information, which can significantly enhance the accuracy and efficiency of identification. (4) The research on deep network architecture that enables the network to have "feedback mechanism" will be the focus of the following research [4].

At present, more widely used programs in the field of computer vision are developed from the starting point of accelerating the speed of image processing [9]. We found that in the whole system architecture can be used, it's very convenient to use the basic methods and functions of, at the same time there are some shortcomings: (1) on the prospects of development in the future, portability can determine the direction of technical development, but relatively widely used in domestic software manufacturers are unable to support this feature. (2) The core algorithm used in most of the software is still developed by foreign researchers, and the core algorithm technology mastered by itself is not
much [2]. (3) Some existing network servers cannot support the development of computer vision technology, and the realization of many functions is limited to the local area. Computer vision needs to process a large amount of pixel data, as well as powerful processor and bandwidth [5].

In view of the above deficiencies, Intel Corporation established OpenCV, which is built in the computer vision libraries of different platforms, with strong versatility and portability, and can be developed on the mainstream operating system. It provides Python, Matlab and other programming languages processing interface; It also supports standard algorithms for pixel processing in different situations.

3. System architecture and validation

Figure 1 shows the architecture diagram of the whole system. We trained the model based on the collected data, obtained the training file of the model using YOLO in the PyTorch environment, and used the model for commodity detection and identification in the actual production environment.

3.1. Preparation of training sets

The training accuracy of object target detection is based on the size of the training set. If the amount of training data is too small, the error of training will become smaller and the error of testing will increase correspondingly, which is called "overfitting". For this purpose, 1000 items were taken as an initial training set, with 12 images of each item from different angles. The training set was augmented by a Python script program, and the original image was increased three times by rotations of 90°, 180°, 270° and horizontal flips. The total number of training sets is 1000*12*5= 60,000.

3.2. Preparation of training set tags

After generating the training set, we also need to generate the LABEL tag for the item. In preparing the training set, barcode labels for each item were collected simultaneously. The 60,000 images are named in commodity codes, again using Python scripts. The result is a TXT file, with each line containing the address of each image and a barcode for each item.

3.3. Preparation of the test suite

The function of the test set is to detect the recognition effect of the generated training set. The test set must not be repeated with the pictures in the training set, otherwise it will violate the requirements of the test. Also in the form of script, the initial pictures are randomly allocated and the test set is generated. The test set and the training set are randomly divided in a 3:7 ratio, which is done through Python's built-in random functions.

3.4. Preparation of test set tags

The test set generated by random division is used for testing. It is also necessary to divide the corresponding labels of the items at the same time. At this time, the division is one-to-one. That is, in the test set, an object has a corresponding lable label, which is stored in the lable_test.txt file. The function of label is a mark that is finally used to classify the identified items. The process of identification is generated by manual labeling.

3.5. The training process

The goal of the training is to improve the mAP (mean Average Precision). The map is a value to measure the accuracy of prediction in object target detection, and it is also the final prediction index. When there are multiple categories of objects to be recognized in the picture to be predicted, a curve of different specifications can be drawn for each category of objects. The value of AP is the area enclosed by the curve on the two-dimensional mAP, while the value of mAP is the area distribution of different curves formed by multiple categories in the picture. After the pictures of the training set are prepared, the images are first identified in the whole region, and the 416*416 images are convolved by CNN to obtain the 13*13 feature map. Each square of the feature map contains corresponding
parameter information and confidence (x,y,h,z,conf). The values of these parameters determine whether there are objects in the region and what kind of objects they belong to.

The final format of the data set: there are 60000 images in the training data set, and 60000 XML files are corresponding to the images. The format of the training set is JPG, which is similar to the format of VOC (Official Training Set). The files annotated with the same naming method and file policy are XML, and the names correspond to the pictures, ranging from 00001.xml to 60000.xml. There are a total of 60,000 pictures, which are randomly divided according to the 7:3 ratio of the Training set and the test set. Among them, 42,000 are used for Training and 18,000 are used for Testing. In all these Training sets, the total category is 1000 kinds, which are 1000 kinds of commodities taken when the data set is just prepared. In which the content of the XML file to the location information of the object, this is our manual annotation files, the inside of the location information is relative information, such as: 0.5 0.6 0.3 0.2, and after yolo network model trained model output are pixels detection information form, as in 480 * 480 images, objects in the image information is highlighted in pixels: 00011 0.0075 336.25684 125.2356 268.6589 165.2547. An object was detected in the file 00011.jpg, with a confidence of 0.0075, at the position 336.25684 125.2356 268.6589 165.2547 of the original image. The result is that the format of the data does not correspond when evaluating the value of the mAP, so you need to look for the corresponding information from the XML file.

Using YOLOv2 to process 1000 items, the content of the file is changed to 1000 barcodes of items, with one barcode per line. During the test of this file, the system will label the product information on the image. That is, after the system boxes the Bounding Box on the image, the text in the upper left corner of the Bounding Box is a line of information in the folder.

The process of training is slow. Considering the time factor, the initial data is 5000 pieces, and the system will have a default iteration process. The default iteration is 45000 times. With the support of I7 processor and GTX2080TI graphics card, it takes three days to train.

3.6. The testing process

The testing process is relatively simple. The preliminary preparation has set up the environment required for the test. It only needs to run the test program after the training generates the weights file. The results after running are shown in Table 1:

| ID | Rps/Img | IOU     | Recall   | Precision |
|----|---------|---------|----------|-----------|
| 101| 1.07    | 82.00%  | 95.50%   | 97.25%    |
| 102| 1.07    | 82.11%  | 95.54%   | 97.27%    |

We define the parameters as follows: ID: iteration number, that is, the frequency of training program to update weights file; Correct: the number of Correct box selections. After processing the data information of a picture, IMGNet will predict the box selections of different objects, compare the Groud Truth of each object with all predicted box selections, and calculate IOU. If the maximum IOU value exceeds the preset threshold, then add one parameter to Correct. RPS/IMG: Region, P Proposals indicate how many frames are predicted per map on average. The value of this parameter is determined by a threshold value. In the YOLO_RECALL function body, the threshold value is set to 0.001. As a result, more boxes will be selected, and some elements that are not objects in the image will be considered objects and selected. However, the benefit of this method is that the value of recall can be increased [2]. The increase of recall value will improve the accuracy of recognition and prevent the object in the image from being unable to be recognized. The Precision value is the ratio.

\[
\text{Precision} = \frac{\text{tp}}{n}
\]

N meaning for (True positives + False positives), that is, the number of network consensus out of the object; TP : True positives, predict the number of correct positions; Recall: The ratio of the number
of correctly identified objects of a certain class to the total number of objects of the same type in the test set. The denominator is True positives + False negatives, which can be understood as the number of a certain type of object.

\[
\text{Recall} = \frac{tp}{tp + fp}
\]

IoU: refers to the coincidence degree between the predicted object position obtained through the network and the original manually marked object position. The value of IoU can be obtained through the set operation of DetectionResult and GroundTruth:

\[
\text{IoU} = \frac{\text{DetectionResult} \cap \text{GroundTruth}}{\text{DetectionResult} \cup \text{GroundTruth}}
\]

3.7. Evaluate performance and improvements
When the threshold is adjusted, the overall recognition rate of the object remains around 98%, which can be improved by adjusting the threshold, as shown in Table 2. By adjusting the threshold, the accuracy of object recognition is higher. The threshold is simply understood as the vertical line moving up and down. If the threshold is small, more images will be predicted successfully, which makes the value of recall larger. The precision-recall curve seeks a relatively balanced value between Precision and recall to obtain the best performance of the system. The changes in Precision and recall values can be observed when the threshold is adjusted. A good classification standard is to identify as many correct objects as possible, and avoid identifying erroneous similar objects, that is, to ensure a certain increase in the recall value, while maintaining a high precision accuracy. Some poor classification ideas will lose the precision accuracy that has been identified for the sake of the recall value, but the precision-recall curve can solve this problem relatively well. Using the classifier precision-recall curve, while ensuring high accuracy, the recall rate can also be maintained at about 40%. When the recall rate reaches 100%, the accuracy rate can be reduced to 50%.

Table 2. Training confidence threshold.

| Retrieval Cutoff | Precision | Recall |
|------------------|-----------|--------|
| Train confidence=0.25 | 98.85% | 20% |
| Train confidence=0.30 | 95.50% | 40% |
| Train confidence=0.35 | 66.34% | 40% |
| Train confidence=0.40 | 73.87% | 60% |
| Train confidence=0.45 | 60.38% | 60% |
| Train confidence=0.50 | 65.65% | 62% |
| Train confidence=0.55 | 55.32% | 80% |

Figure 2. Test results.
4. Conclusion
Self-help shopping in the supermarket system is the core content in commodity target detection problem, we use a degree of knowledge of computer vision, YOLO with popular object recognition algorithms, and key to YOLO introduced in this article, at the same time, the environment of object detection target recognition system, the system using the Python language. Python as a high-level language makes it relatively slow to run programs on it, but in the age of rapid advances in computer power, we are increasingly looking for a language that is easier to read and modify. In the system, another important aspect is the real ground practical science and technology, computer vision technology as the core of artificial intelligence field, combined with NLP (natural language processing) will be more closely together, will be in industry produces more efficient than human productivity, augur well for the future will be more now appears to be "stable" work will lose their jobs.

At present, computer vision is limited by the speed problem. Under the condition of ensuring the recognition accuracy, the recognition speed is still difficult to improve. In the future development, the recognition speed will be greatly improved, and will be used to capture human facial expression changes, infer human emotions, through the imported psychological model, the computer will have more acute observation power. In the future, with China's attention to information technology, information technology will continue to play a greater role in promoting the development of all walks of life in the whole society, and the trend of the development of informationization and networking in traditional industries will become more and more obvious. The informationization of products and life will prove that informationization plays an irreplaceable role in promoting China's economic development. Medical informatization will also follow the footsteps of mass informatization development more rapidly.

Acknowledgements
In the process of writing the paper, Professor Liu Yu and Professor Liu Dongchang gave me a lot of guidance. At the same time, I am very grateful to Li Yuge from Renmin University of China, Cui Shuhao from Tsinghua University, and Mr. Wang from Institute of Computing Technology, Chinese Academy of Sciences. They fully explained the most popular image recognition technology to me, which played an important role in the preparation of my paper. At the same time, Wang Jing gave me great encouragement and guidance in Huawei's Beijing Research Institute where I worked as an intern, and I also got a lot of inspiration from the process of consulting engineers in Baidu's Technology Department. Finally, I would like to express my gratitude to Mou Jia, my classmate in the lab, who gave me a lot of help in writing the paper.

References
[1] You only look once:unified,real-time object detection.Redmon J,Divvala S,Girshick R,etal.Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.2016.
[2] A closer look at Faster R-CNN for vehicle detection.Fan Q,Brown L,Smith J.Intelligent Vehicles Symposium (IV).2016
[3] YOLO9000:Better,Faster,Stronger.J Redmon,A Farhad.Computer Vision and Pattern Recognition (cs.CV).2016
[4] A Lightweight YOLOv2:A Binarized CNN with A Parallel Support Vector Regression for an FPGA.H Nakahara,H Yonekawa,T Fujii,S Sato.Acm/sigda International Symposium.2018
[5] Xu K, Ba J, Kiros R,et al. Show, attend and tell: Neural image caption generation with visual attention[C]/International conference on machine learning. 2015: 2048-2057.
[6] WEI W, JIA Q, CHEN G. Real-time facial expression recognition for affective computing based on Kinect[C]/ Industrial Electronics & Applications. IEEE,2016:161-165.
[7] C. Feichtenhofer, A. Pinz, and R. Wildes. Spatiotemporal residual networks for video action recognition. In NIPS, 2016. 2, 3
[8] J. Redmon and A. Farhadi. Yolov3: An incremental improvement. arXiv, 2018. 4
[9] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar. ‘Focal loss for dense object detection.
[10] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016.

[11] A. Shrivastava, R. Sukthankar, J. Malik, and A. Gupta. Beyond skip connections: Top-down modulation for object detection. arXiv preprint arXiv:1612.06851, 2016.