Research on Automatic Inspection of Product Production Based on Computer Vision

Yangjia Zhou

Shanghai Maritime University, Shanghai 200135, China

crazy_zhou10@163.com

Keywords: Computer vision, DL, Automatic product detection

Abstract: The visual information contained in the product is complex, and many kinds of information overlap with each other, so a single category label can't fully describe it. With the great development of AI (Artificial Intelligence) theory and application, the accuracy and stability of image recognition algorithm based on DL (Deep Learning) have been greatly improved, and it has gradually met the needs of various visual application scenarios. In this paper, we use computer vision and DL technology to study how to automatically identify products in images. In chapter 1, we propose an efficient and concise target detection network: DPF PN-net (dual path fusion feature pyramid constructive network). A DPFM(Dual Path Fusion Module) is used to circularly perform two-way feature fusion on the three-layer feature map to increase the reuse of fused features, and a product qualification detection method based on improved Mask R-CNN is proposed. Experimental results show that the improved Mask R-CNN algorithm in this paper has high recognition rate, and the detection speed is better than that of the original Mask R-CNN algorithm.

1. Introduction

At present, the shape and position of products are detected by manual, static and several single-point measurements, which can no longer meet the requirements of today's industrial production in terms of detection accuracy, detection efficiency and labor intensity. The classification of human eyes and state tracking are not only inefficient, but also prone to errors. Therefore, photoelectric scanning and image recognition technologies are adopted to meet the requirements of automatic classification after automatic recognition [1]. Computer is used to control the operation of the whole system and the setting of parameters, so it has the characteristics of complete functions, good portability and convenient regulation.

At present, bar code identification technology is widely used to identify commodities in the production, storage, sales and after-sales service of commodities. Radio frequency identification (RFID) is widely used in emerging unmanned convenience stores, unmanned containers and other scenes. Literature [2] In this study, the image classification scheme based on DL (Deep Learning) AlexNet won the competition championship in one fell swoop, and the accuracy rate far exceeded the second place. Since then, deep learning has become the mainstream method to solve various tasks in the field of computer vision. The three basic tasks of computer vision are classification, detection and segmentation. Literature [3] The image classification scheme based on DL model AlexNet won the competition in one fell swoop, and the error rate was nearly 10% lower than that of the second place. Since then, it has aroused great concern and in-depth research of deep learning in academia and industry.

Considering the cost and efficiency, starting from the requirements of reliability, accuracy and real-time, this paper uses traditional computer vision processing technology and DL technology to dynamically detect and identify the products on the conveyor belt, and then completes the product settlement process. In this paper, the product settlement on the conveyor belt in production is taken as the application scenario, and the traditional computer vision processing technology is used as an auxiliary means to study the product detection and identification method based on DL, which is also of reference value to other detection and identification problems.

  DOI: 10.25236/iiiece.2022.033

2. Overview of DL

As can be seen from Figure 1, DL is a network structure with many hidden layers developed from artificial neural networks. Multilayer perceptron is a model representing the nonlinear mapping between input and output vectors. It is the superposition of many simple nonlinear transfer functions, which enables the multilayer perceptron to approximate the nonlinear function [4]. The output of the node is scaled according to the connection weight and used as the input of the next network node, which means the direction of information processing. Therefore, multilayer perceptron is also called feedforward neural network.

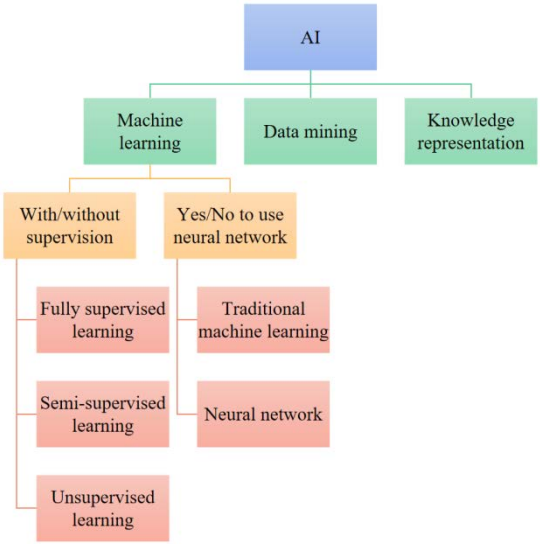

Fig.1 Research Direction and Method of Artificial Intelligence

The structure of multilayer perceptron is variable, usually composed of multilayer neurons, and its structure is shown in Figure 2. It combines low-level features to form abstract high-level features, which are used to represent attribute categories or features, so as to represent data by features.

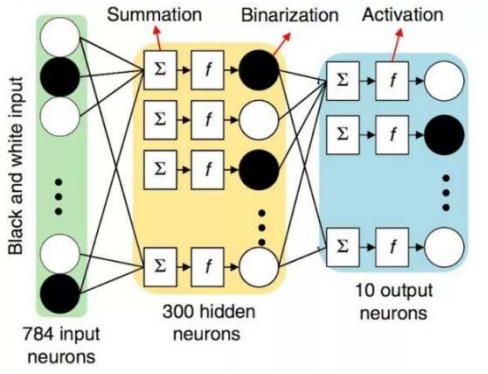

Fig.2 Multilayer Perceptron

As far as the specific research content of DL is concerned, it mainly involves three types of methods [5]:

(1)Neural network system based on convolution operation, namely CNN (Convolutional Neural Network);

(2) Self-coding neural networks based on multi-layer neurons, namely self-coding and sparse coding;

(3) Pre-training with multi-layer self-coding neural network, and further optimizing the DBN (Deep Belief Network) of neural network weights through identification information;

According to the three methods mentioned above, this paper will use CNN to classify the quality of products. CNN is inspired by the structure of vision system, which makes the research of computer vision develop rapidly. CNN is one of the representative algorithms of DL, which is

different from traditional shallow learning, and its difference lies in [6]:

(1)Emphasize the depth of the network structure, usually with 5, 6 or even 100 layers of hidden nodes;

(2) Clear new ideas of feature learning. Through the nonlinear superposition of data features layer by layer, the feature representation of data in the original space is transformed into a new feature representation, which makes the classification or prediction more accurate. Compared with traditional feature extraction methods, using massive data to learn features in data can extract more useful information from data.

By designing the number of neuron computing nodes and the hierarchical structure of multi-layer operation, selecting the appropriate number of input nodes and output nodes, and by learning and optimizing the network, the functional relationship from input to output is established. Although the functional relationship between input and output cannot be found 100%, it can approach the correlation between real data as much as possible. Through repeated training of the model, we can get a network model with excellent prediction or recognition effect, which can meet our demand for automation of complex transaction processing.

3. Automatic Product Detection

3.1 Image Mosaic and Target Location

Considering the accuracy of target location in practical application, RGB-D sensor is used in product detection and identification system, which can not only collect RGB color information, but also obtain the distance from each point on the surface of the object to the sensor, that is, the depth information. In the system, the RGB-D sensor is responsible for collecting the product images on the conveyor belt. When the conveyor belt is running, the RGB-D time series images will be obtained.

ORB (oriented fast and rotated brief) algorithm is a common feature point extraction algorithm, which is divided into feature point detection and feature point description [7]. Feature point detection is improved from FAST algorithm, which detects the pixel value of a circle around a candidate feature point in a fast way. If there are enough pixel points in the field around the candidate point and the gray value of the candidate point is sufficiently different, the candidate point is considered as a feature point.

The key of image mosaic is to determine the geometric relationship between point coordinates in different images and build a coordinate transformation model. The geometric relationship between images can be expressed by a global homography matrix under the condition that the photographed objects are on the same plane or the spatial position of the photographed viewpoint does not change.

Find the homography matrix H according to the matching point pairs to minimize the fitting error $E(H)$:

$$\min_{H} E(H) = \min_{H} \sum_{(P,Q)\in C} \left\| H(P) - Q \right\|_2 \quad (1)$$

Where, $H(P)$ is to convert the coordinate point P in image I_i into image I_j by using homography matrix.

Because each pair of feature points can construct two linear equations, it takes at least four pairs of feature points to determine the homography matrix with eight degrees of freedom. In this section, RANSAC (RANdom SAmple Consensus) algorithm is used to solve the homography matrix [8].

Firstly, four pairs of feature points are randomly sampled to obtain a homography matrix H; Then, H is applied to all feature point pairs, and the fitting error is calculated. If the error is greater than the threshold value, the pair of feature points can not be fitted. Repeat for many times to get a H^* that can fit the most number of feature point pairs.

Finally, the perspective transformation of RGB time-series images and registered depth time-series images is carried out by using H^* to unify the pixel coordinates of different images.

Coordinate transformation can unify the pixel coordinates of two images into a coordinate system, and the images obtained by directly averaging the overlapping areas will have a very obvious transition phenomenon in the overlapping areas of two images. In this section, the weighted average method is used to deal with the overlapping areas of two images, and the formula is:

$$\bar{v}_{ij}^{3}=\begin{cases} v_{ij}^{1}, v_{ij}^{3}=0 \\ \alpha v_{ij}^{1}+(1-\alpha)v_{ij}^{2}, v_{ij}^{3}\neq 0 \end{cases} \quad (2)$$

Where i, j represents the row number and column number respectively, v_{ij}^{1}, v_{ij}^{2} is the pixel value in the overlapping area in the first and second images after coordinate transformation respectively, and v_{ij}^{3} is the pixel value after direct average processing of the overlapping area.

$\bar{v}_{ij}^{3}$ is the pixel value after weighted average; $\alpha=1-\frac{j-left}{right-left}$ represents the fusion weight, where $right, lef$ represents the column numbers of the left and right borders of the overlapping area, respectively. This method enables the overlapping area to be gently transferred from one image to another.

3.2 Product Inspection Based on DPFPN-Net

During the training and testing of the network, there will be a large number of shifts in the input samples, which will change the pooled output, and the same input will produce different characteristic information, which will make the network unstable and difficult to converge and reduce the detection accuracy. In order to solve the problem that one-way fusion is greatly affected by pooling offset and increase the feature reuse rate in the fusion mechanism, this chapter proposes an efficient and concise target detection network: DPF PN-net (Dual Path Fusion Feature Pyramid Convolutive Network). A kind of DPFM(Dual Path Fusion Module) is used to circularly perform bidirectional feature fusion on the three-layer feature map, which increases the reuse of fused features, enriches semantic and location information, and suppresses the influence of pool migration.

DPFPN-Net uses the optimized MWI-DenseNet as the convolution backbone network, selects the feature maps of some layers in the backbone network as the original feature pyramid, and uses DPFM to analyze a feature map in the original feature pyramid from top to bottom and bottom to top [9]. The feature maps of the previous layer and the latter layer are fused, each DPFM fuses three feature maps, and multiple DPFM loops are fused in turn, which brings rich semantic information and accurate positioning information to each prediction layer, which is beneficial to improve the detection accuracy of various scale objects, and at the same time, the feature maps are circularly reused, which reduces the calculation amount of feature fusion.

As shown in Figure 3, DPFM is different from FPN(Feature Pyramid Network), and each DPFM module fuses the features of the current layer, the previous layer and the next layer in the original feature pyramid, corresponding to the layer shown by *(i)-th layer,(i-1)-th layer, (i+1)-th layer* in Figure 3.

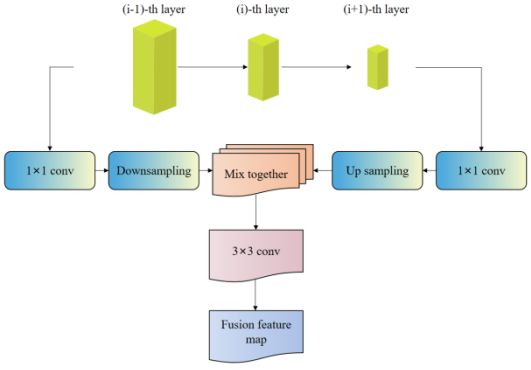

Fig.3 Dual-Path Feature Fusion Module

Fig. 4 shows the construction of feature pyramid using 3 DPFM loop fusion. The DPFM corresponding to the 4-thlayer fuses the feature maps of the current layer 4-th layer, the previous layer 3-th layer and the next layer 5-th layer, and the fusion result is used as the input of the top-down fusion path of the DPFM corresponding to the 3-th layer. The 2-th layer is used as the input of the bottom-up fusion path, and the fusion operation is repeated again.

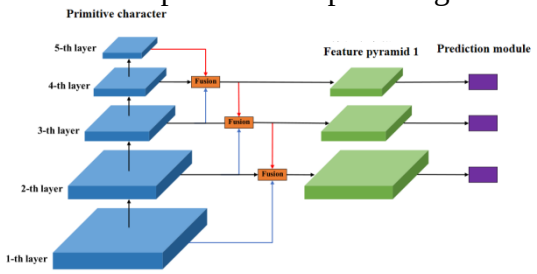

Fig.4 Basic Dual-Path Fusion Feature Pyramid Detection Network

Experiments show that this cyclic multiplexing bidirectional feature fusion pyramid network can effectively improve the detection accuracy, especially for small objects. At the same time, compared with many mainstream feature fusion pyramid networks at present, this fusion mechanism has less calculation amount of parameters, and the network is more concise and efficient.

In order to further improve the semantic representation ability of feature pyramid, bring more accurate position information to the deep feature layer, and further solve the problem of feature loss and change caused by pooling, a recombination and fusion module is proposed.

Split the input $(i\text{-}1)\text{-}th\ layer$ feature map. The splitting method is to divide the feature map into four sub-blocks on average according to the length and width according to the Passthrough method. Each sub-block has the same length and width as $(i\text{-}1)\text{-}th\ layer$, and the four blocks are spliced in the channel dimension in a fixed order. Results The stitching $(i\text{-}1)\text{-}th\ layer$ was recombined and spliced, and then 1×1 convolution operation was performed.

3.3 Improved Mask R-CNN

Mask R-CNN mainly includes feature extraction network, FPN(Feature Pyramid Networks), RPN(Region Proposal Network) and Fast-RCNN+mask. Among them, feature extraction network and FPN are the backbone networks, which are used to generate feature maps, RPN to generate suggestion areas, Fast-RCNN to classify and regress, and mask to classify and regress targets at pixel level.

In Mask R-CNN, the implementation of feature extraction network uses ResNet50, which has excellent performance [10]. Compared with other CNN, ResNet50 residual learning mechanism enables deep network to avoid degradation. Degeneration refers to the decline of network accuracy with the increase of network depth.

FPN is for better feature fusion. Common networks generally use the feature map of the last layer. Although the feature map of this layer has strong semantics, its position and resolution are low, and the detection effect for small objects is not good. However, FPN network fuses the features from the bottom layer to the top layer, making full use of the features extracted in each stage, and greatly improving the detection performance of the model for small target objects [11].

The structure of RPN was proposed in Faster R-CNN, and the sliding window task of serial processing was changed into the anchor task of parallel processing by anchor point, which greatly accelerated the processing speed [12].

The full connection layer of Mask R-CNN model also encounters such huge parameters, and more parameters will affect the speed of model training and product inspection. Therefore, this paper proposes to use the global average pooling layer to replace part of the fully connected layer, which can save a lot of resources. In the Mask R-CNN classification and regression module, as shown in Figure 5.

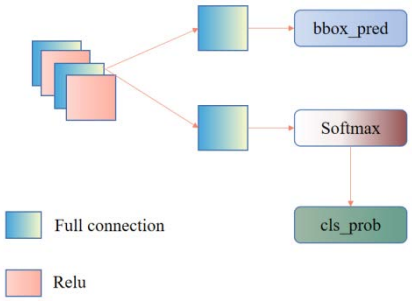

Fig.5 Mask R-CNN Classification and Regression Module Structure Diagram

Use a global average pool layer to replace the full roll-up layer in front of the classification module. This method greatly reduces the model parameters, avoids the occurrence of over-fitting, and reduces the storage space occupied by the model. Global average pooling operation makes the output features have global receptive field, which can make use of global information.

The number of parameters is an important factor that affects the training time and recognition speed. According to the analysis of feature extraction network in Mask R-CNN, the convolution kernel used by Conv1 in the model is 7×7, and the elements in each convolution kernel need a weight, and the convolution kernel needs a total of 49 weights.

In this paper, it is proposed that Conv1 uses convolution kernels with the size of 3×3 and the step size is set to 1, and uses three small convolution kernels instead of convolution kernels with the size of 7×7, so that the receptive fields of the obtained feature maps are the same, and the total number of weights is reduced by 22, at the same time, the nonlinearity of hidden layer features is enhanced, as shown in Figure 6.

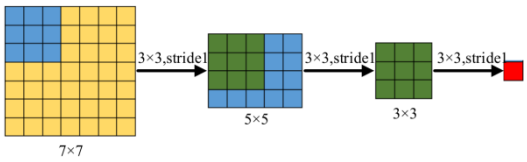

Fig.6 Schematic Diagram of 3×3 Convolution Kernel Replacing 7×7 Convolution Kernel

Due to the increase of nonlinear superposition and the decrease of the number of parameters, the training time of the model can be shortened and the recognition speed can be improved.

4. Experimental Results and Analysis

This experiment is carried out on the Linux operating system Ubuntu, and the experimental program is written in python language based on TensorFlow machine learning development framework. Fig. 7 shows the parameters and reasoning time of each model.

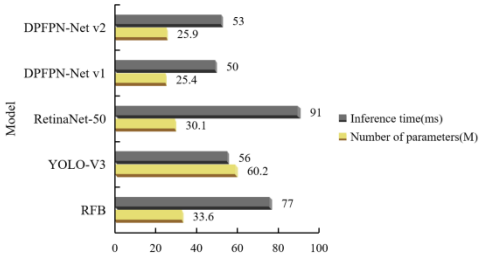

Fig.7 Parameter Quantity and Reasoning Time of Each Model

As can be seen from Figure 7, large-size targets have higher confidence and more accurate detection frame than small-size targets. The target in simple background picture has higher confidence than the target in complex background, and the detection box is more accurate. At the same time, it can be seen that occlusion, partial imperfection and other problems will have some influence on confidence and bounding box prediction, but the model is robust, occlusion imperfection will not lead to missed detection and misjudgment of the model within a certain range,

and the output prediction result is still good.

By analyzing the experimental results, it can be seen that the product detection model DPFPN-Net v2 designed in this section has high detection accuracy, effectively improves the detection accuracy of small objects, can adapt to the detection of targets of various scales, and at the same time, the reasoning time of the model is less, which can meet the real-time requirements during deployment.

According to fig. 8, the detection effect of Mask R-CNN is better than that of Faster R-CNN. Compared with the original Mask R-CNN detection effect, the improved model in this paper improves the average accuracy mAP by 4.7 percentage points and the detection speed by about 0.09 seconds.

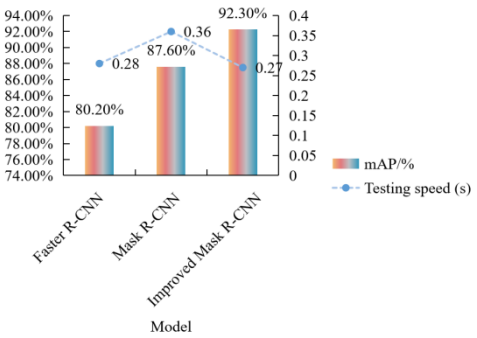

Fig.8 Comparison of Experimental Results of the Model

In this paper, the improvement of the detection accuracy of Mask R-CNN is mainly due to the introduction of nonlinear layer, which improves the ability of feature extraction and feature expression of the model, and the reduction of model parameters makes the detection speed improved.

5. Conclusion

Product inspection is an important part of product circulation. At present, there are some problems in the process of product inspection, such as low efficiency or high cost, which affects people's purchasing experience in stores or production and causes certain sales losses. This paper aims to realize a product detection and identification system by using computer vision and DL technology, and apply it to the product detection process. The feature pyramids are fused again by RFM, and finally a three-feature pyramid convolution network DPFPN-Net v2 is formed. The experimental results show that the proposed detection model has high detection accuracy, effectively improves the detection accuracy of small objects, can adapt to multi-scale target detection, and has less reasoning time. The improved Mask R-CNN algorithm has high recognition rate, and the detection speed is better than that of the original Mask R-CNN.

References

[1] Gao Q, Liu W, Li D, et al. Research and Implementation of the Roll Position Automatic Adjustment System Based on Roller Parameters Prediction. Journal of Advanced Manufacturing Systems, vol. 18, no. 2, pp. 273-292, 2019.

[2] Xiang Yu, Gao Jianpo, Ren Zhenxing. Research on automatic detection system of nonel tube based on machine vision. Electronic design engineering, vol. 027, no. 007, pp. 56-60, 2019.

[3] Yu X, Wang Z, Wang Y, et al. Edge Detection of Agricultural Products Based on Morphologically Improved Canny Algorithm. Mathematical Problems in Engineering, vol. 2021, no. 3, pp. 1-10, 2021.

[4] Li B, Yang J, Zeng X, et al. Automatic Gauge Detection via Geometric Fitting for Safety Inspection. IEEE Access, no. 99, pp. 1-1, 2019.

[5] Ya, Li Qingnan, Liu Xiaochen. Automated visual inspection of target parts based on deep learning. message communication, vol. 000, no. 002, pp. 50-53, 2019.

[6] Duan L, Yang K, Lang R. Research on Automatic Recognition of Casting Defects Based on Deep Learning. IEEE Access, no. 99, pp. 1-1, 2020.

[7] Li Changbai, Yang Jie, Huang Zhitao, etc. Automatic Modulation Recognition of Communication Signal Based on Deep Learning. Space electronic technology, vol. 016, no. 001, pp. 49-54,74, 2019.

[8] Y Kumar, Sheoran M, Jajoo G, et al. Automatic Modulation Classification based on Constellation Density using Deep Learning. IEEE Communications Letters, no. 99, pp. 1-1, 2020.

[9] Li L, Dong Z, Yang T, et al. Deep learning based automatic monitoring method for grain quantity change in warehouse using semantic segmentation. IEEE Transactions on Instrumentation and Measurement, no. 99, pp. 1-1, 2021.

[10] Xie Q, Li D, Xu J, et al. Automatic Detection and Classification of Sewer Defects via Hierarchical Deep Learning. IEEE Transactions on Automation Science and Engineering, vol. 16, no. 4, pp. 1836-1847, 2019.

[11] He S, Chen L, Zhang S, et al. Automatic Recognition of Traffic Signs Based on Visual Inspection. IEEE Access, no. 99, pp. 1-1, 2021.

[12] Sampedro C, Rodriguez-Vazquez J, Rodriguez-Ramos A, et al. Deep Learning-Based System for Automatic Recognition and Diagnosis of Electrical Insulator Strings. IEEE Access, no. 99, pp. 1-1, 2019.