Intelligent monitoring method of tridimensional storage system based on deep learning

Mingzhou Liu1 · Xin Xu1 · Xiaoqiao Wang1 · Qiannan Jiang1 · Conghu Liu2,3

Received: 13 January 2022 / Accepted: 2 May 2022 / Published online: 19 May 2022
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2022

Abstract
Growing international trade requires more flexible warehouse management to match it. In order to achieve more effective warehouse management efficiency, a shelf status–detection method based on deep learning is proposed. Firstly, the image acquisition of a multi-level shelf containing multiple bays is performed under different time and lighting conditions. Due to the difference in image characteristics between the bottom shelf on the ground and the upper shelf on the non-ground level, the collected images were divided into two groups: floor images and shelf images; and the warehouse status recognition was performed on the two groups separately. The two sets of images are cropped and center projection transformed separately to obtain the region of interest. On this basis, the improved residual network model is used to construct different depot detection models for the two sets of images, respectively, and the above algorithm is verified by actual measurements. In this paper, 102,614 images of 3246 depots with different states of non-ground layer, and 27,903 images of ground layer are collected. They are divided into training set and test set according to the ratio of 4:1, and the accuracy of training set is 99.6%, and the accuracy of test set is 99.3%. The experimental outcomes provide a theoretical method and technical support for the intelligent warehouse system management.

Keywords  Shelf’s status detection · Deep learning · Warehouse management · Logistics · Energy consumption

Introduction
Since the outbreak of COVID-19, global trade competition has become increasingly fierce. The operation speed of the supply chain needs to be accelerated, which requires a more flexible and intelligent warehousing system. Inventory management is increasingly demanding in terms of accuracy and real-time, and the timeliness of material distribution is becoming more and more important.

The logistics industry, as a result, has been regarded as a new accelerator of China’s economic prosperity. However, a series of environmental issues such as energy consumption, air pollution, carbon emissions, land redundancy and waste, and traffic congestion, as well as subsequent social and economic issues have challenged the development of logistics industry significantly (Dai and Gao 2016; Xie et al. 2016; Fan et al. 2017). When considering total carbon emissions from 1997 to 2017, direct and indirect CO2 emissions in China’s logistics industry roughly followed the same upward trend, but with large differences in the total amount (Liu et al. 2021). Moreover, carbon emissions in storage, transportation, logistics, and other fields are seriously underestimated (Lang 2018; Lang et.al 2021). Managing a logistics system involves several related activities, i.e., warehousing, inventory handling, information services, and transportation, and any decisions may influence a large number of stakeholders in either positive or negative ways (Huang et al. 2021).

The effectiveness and sustainability of a logistics system determine the long-term competitiveness and the success of an enterprise. Therefore, new methods are investigated by both academia and industrial practitioners to improve the economic, environmental, and social sustainability
of logistics activities (Xu et al. 2021). Logistics sustainability is increasingly becoming a central focus of businesses, when most societies are aware of the influence of industry on both the environment and human health. To address the drawbacks of the way, logistics systems have been designed, a new logistics system called Physical Internet has been proposed (Fergani et al. 2020). Therefore, in a specific logistics process, it is necessary to use new technologies to increase efficiency and reduce carbon emissions and energy consumption.

The energy and resource consumption in traditional warehousing mainly lies in the up and down storage of materials, the flow of materials, and the consumption of warehousing personnel. Due to the relatively backward management, the additional resources and energy consumption caused by the flow of materials becomes high (Xu et al. 2021). Inventory items are always in a constant dynamic of incoming, outgoing, and inevitably make mistakes in the process of operation. The errors generated during the operation will accumulate over a period of time, which will make the data reflected in the inventory information and the actual data not match, and some items will change in quantity and quality due to long storage time and improper storage. In order to truly reflect and grasp the quantity and quality of inventory items and to carry out effective control, it is necessary to conduct inventory operations. Warehouse inventory is an important part of warehouse management, but in general, traditional warehouse is still manual inventory count.

Manual inventory is less efficient; the general situation of the warehouse in 1-2 weeks inventory once; the entire warehouse thorough inventory is generally once a month, or even once a quarter; manual inventory is also prone to errors and omissions; inventory status is not updated (Senthil Nathan et al. 2018; An et al. 2021; Tavana et al. 2021; Yesenia and Micheal 2019). So, the daily inbound and outbound activities, warehouse managers are difficult to each outbound and inbound warehouse to have a comprehensive understanding of the overall warehouse status. Each time the warehouse position changes, warehouse managers need to go to the location of the warehouse position to confirm the status of the warehouse position. The current technologies for depot status identification are QR code (J Egger et al. 2018), magnetic card magnetic stripe (Liukkonen and Tsai 2016), RFID (Ding et al. 2018; Liu et al. 2019; Liu et al. 2015), Bluetooth (Sangrok et al. 2019), infrared (Chen and Lei 2019), etc. QR code and magnetic card magnetic stripe need handheld terminal equipment for close range identification. RFID, Bluetooth, infrared, and other library level marking methods also have the disadvantages of high cost, susceptibility to interference, and poor stability (Baesso and Oliveira 2018; Pablo et al. 2016).

At present, computer vision technology has made great progress, and the detection technology based on vision has been well applied in many industrial scenarios (such as defect detection, product model identification, and target positioning). The problem of location state detection based on vision is essentially a classification problem. The classification algorithm can be divided into two categories, namely, the supervised classification algorithm and the unsupervised classification algorithm. In the first classification algorithm, the training samples are firstly collected and labeled manually and then sent to the classification model for training. Finally, the trained model is used to classify the test samples. The second kind of classification algorithm does not need to label the test samples, but directly input the training samples into the model to complete the self-classification, and then use the trained model into the test samples (Lecun et al. 2015).

The commonly used supervised classification algorithms are: traditional methods such as SVM, decision trees, random forests, neural networks, and deep learning network methods (Alexander et al. 2017). When detecting the warehouse state based on the aforementioned traditional supervised classification algorithms, experts are required to extract low-dimensional features from the images and use them as inputs to the model. However, it is difficult to extract features from the images that can characterize the shelf state because the warehouse images collected in real application scenarios are affected by the type of goods, shelf color, lighting changes, and shooting angles. The unsupervised learning algorithms are: principal component analysis (PCA) (Assia et al. 2018), traditional clustering methods (Hedjam et al. 2021), self-organizing mapping (SOM) (Chen and Huang 2020), deep confidence networks (DBN) (Zhou and Zhang 2021), deep Boltzmann machines (DRBM) (Tao 2021), self-encoder (AE) (Tang et al. 2021), and generative adversarial networks (GAN) (Huang et al. 2015). The first three unsupervised learning algorithms mentioned above have the same drawbacks as traditional supervised classification algorithms and are therefore not applicable to library status detection. When other network-based unsupervised learning algorithms are used for library state detection, the images are different for each library when it is idle, so it is necessary to build a classification model for each library, which makes the training process more complicated and time-consuming. If the whole image is input into the traditional supervised classification algorithm without dimensionality reduction, it will make the model difficult to train and explode in dimensionality and cannot achieve effective classification. The deep learning network solves the above problems well and can classify the whole image well when it is used as the input of the model. Next, we discuss the innovation of this paper in terms of both hardware layout and algorithm of the library status detection system.
Hardware layout of the detection system

Asaoka T. et al. (Asaoka et al. 2018) developed a book stacking method–detection system based on deep learning using books on a bookshelf as the target object. The system captures the front view of the books from the front of the shelf by a manual handheld mobile camera. Rennie et al. (Rennie et al. 2016) proposed an improved RGBD object dataset warehouse detection and object state estimation system, where the data collection is done by a vision sensor fixed on the robot arm, and the acquisition of target images on different shelves is done by moving the robot arm. Tianjian et al. (Tianjian et al. 2019) designed a system for automatic identification and positioning of pallets in a warehouse, which identifies and locates the pallets in front of a forklift by means of a camera mounted on the front of the forklift. Chatpreecha et al. (Chatpreecha and Keatmanee 2018) constructed a vision-based warehouse state monitoring system that is able to identify whether a warehouse position is in a full, half-stocked, and no-stocked state. The camera of this system is fixed directly opposite to the shelf. However, the sensor layout method approach of the above reference is not applicable to the stereo warehouse scenario studied in this paper.

General three-dimensional warehouse has the following characteristics: large footprint, high height; the distance between the two rows of shelves is large, the number of storage spaces, and the need for frequent up and down the goods.

The way of capturing pictures of the warehouse positions by handheld cameras in the literature (Asaoka et al. 2018) cannot achieve the effect of saving manpower on the one hand, and the upper levels of the stereo warehouse are high and inconvenient to capture manually. The way of data acquisition by fixing the camera on a fixed robot or mobile forklift in the literature (Rennie et al. 2016) and (Tianjian et al. 2019) increases the hardware investment cost, interferes with the logistics vehicle of the goods, and the data collection will take a lot of time, which in turn makes the whole system less time-efficient. If, as in the literature (Chatpreecha and Keatmanee 2018), the camera is placed on the opposite side of the shelf to be inspected, although the front view of the goods position can be collected, it will affect the loading operation as well as the discharging operation causing an impact. If the camera is mounted on the walls at both ends of the warehouse, the view is too partial to capture the full view of the status of the goods level, although it captures the impact operations. Unlike the above two locations, mounting the camera on the top of the warehouse and observing the status of the storage position from top to bottom not only ensures that it does not affect the operation, but also captures the ideal picture.

Therefore, considering the economy and practicality, this paper installs the network dome camera at the top of the warehouse 3 m downward; combining with the above-mentioned characteristics this place is a suitable location for the camera.

Target classification algorithm

As mentioned above deep learning–based classification and detection algorithms are currently the most excellent supervised learning algorithms, and they have been widely used in other fields. For example, Stephen K et al. (Stephen et al. 2021) proposed an improved AlexNet method for vehicle detection and classification in real road environments, and experiments showed that the method has better performance in vehicle detection and car classification. Tianjian L. et al. (Tianjian et al. 2019) proposed a CNN-based algorithm for warehouse pallet recognition, and experimental results showed that the method has a high accuracy and the improved pallet detection algorithm achieved a detection rate of 92.7% and a test rate of 42 frames/s, which can meet the efficiency and accuracy requirements of pallet detection when using TITAN X GPU for practical applications. Zhao K. (Zhao et al. 2020) proposed a ResNet-50 based breast abnormality diagnosis system, and in order to improve the proposed model created a new data enhancement framework, resulting in an overall accuracy of 95.74% for the model; thus, the method is effective in classifying breast abnormalities. Mohamed L et al. (Mohamed et al. 2020) firstly based on ResNet-50 on yolov2 for medical mask–detection algorithm; the results showed that the average accuracy was 81% compared to other supervised classification algorithms; this algorithm obtained higher accuracy and precision. Gebhardt C et al. (Gebhardt et al. 2020) used a simplified residual neural network (SimResNet) to predict fatigue strength from metallographic data, and the experimental results showed that the application of SimResNet was able to accurately predict the fatigue strength obtained from microstructure ansatz analysis of ductile iron. Pan T S et al. (Pan et al. 2020) based on a modified residual neural network (ResNet) for underwater target detection, and the results showed that the recognition rate of the method was 96.5% (mAP). Gu X et al. (Gu and Qu 2020) proposed a face recognition algorithm based on ResNet50 improved reID model, which adds a linear layer, a batch parametric layer, and a ReLU layer before the classifier, and the improved model can achieve good real-time. Hao Zheng et al. (Hao Zheng et al. 2022) proposed a structure optimization method, which replaces VGG16 in fast RCNN with ResNet to make it suitable for small target recognition in complex background.

From the above literature, it is clear that the deep learning–based target detection algorithms not only have high detection accuracy and strong resistance to environmental
interference, but also can meet the timeliness requirements in practical applications. Among them, the detection accuracy of ResNet-50 is the highest. In summary, in order to make the detection system more robust and time-efficient, this paper constructs a new three-dimensional warehouse cargo level status detection system. First, the camera is arranged in the center of the shelf without obstruction and interference to collect the warehouse level image and establish the corresponding database; then the warehouse level area under each camera position is manually cropped from the image, and the center projection transformation is performed; finally, the improved Res-Net model is trained using the labeled training set, and the warehouse level status detection is implemented based on the model. The experimental results show that the system still has high detection accuracy under different lighting conditions, different cargo types, and shooting angles.

Our research provides both theoretical contributions and practical implications. The theoretical contributions are as follows. In this paper, we study the fundamental issues of smart warehousing. We build a deep learning model for intelligent storage systems to achieve quantitative evaluation of the storage status of intelligent storage systems, which in turn provides a comprehensive measure of the operational capability of storage systems in terms of economic efficiency and energy. Its practical significance is as follows. This paper can help production managers analyze the characteristics of the warehousing system from the perspective of the depot status, tell managers how to evaluate the eco-efficiency of the intelligent warehousing system, and improve the logistics scheduling level and core competitiveness of the warehousing enterprise through data-driven. It helps to provide the sustainability of the warehousing system.

Materials

Hardware architecture and inventory status analysis

DS-2PT7D40IW-DE (8–32 mm) is used as the model of net ball machine. Parameter configuration: color 0.002 lx @ (f1. 6, agcon); black and white 0.0002 lx @ (f1. 6, agcon); 0 lx with IR; AGC — automatic gain control; 0 Iluminance of IR integrated camera; main stream: 50 Hz: 50 FPS (1920×1080), 60 Hz: 60 fps (1920×1080). Video compression h.265/h.264/mjpeg; infrared irradiation distance: 50 m; vertical preset point speed: 200°/s. Horizontal preset point speed: 350°/s. The hardware layout of this paper is shown in Fig. 1.

According to different needs (such as rotation training cycle, recognition accuracy, camera parameters), a camera shooting angle can correspond to multiple locations, can also correspond to a location. In order to fully obtain the location information, this paper uses a camera shooting angle corresponding to a location state. The comparison chart of goods and no goods is shown in Fig. 2.

Although there are only two kinds of stock and no stock, the image features of non-ground layer (i.e., shelf layer) and ground layer (i.e., no shelf layer) are different. If the above two categories are treated as one, the accuracy of the classifier will be affected. Therefore, in this paper, states are divided into four categories: no goods on the ground floor (a), goods on the ground floor (b), no goods on the shelf floor (c), goods on the shelf floor (d).

At the beginning of this paper, the goods and non-goods of all the warehouses were classified and labeled, but the recognition rate has been around 90%. After repeated research, it was found that there are different characteristics between the ground level and shelf-level warehouses, as shown in Fig. 3. The yellow reflective warning line is pasted on the ground, and the goods cannot press this line. In the picture of no goods in the first line and goods in the second line in Fig. 4, there is a yellow reflective warning line; we can see this yellow reflective cordon. However, if the goods are placed on the shelf level, the goods will definitely block the shelf crossbeam from the camera's perspective, so the pictures of the shelf level are like lines 3 and 4 in Fig. 4. You can see that the first three lines have the same characteristics (the bar at the bottom). So, in this paper, the images of ground level and shelf level are trained separately. From the final results, the recognition rate is greatly improved. The red line indicates the main identification area.

Method

Firstly, the image acquisition of a multi-level shelf containing multiple bays is performed under different time and lighting conditions. Then, the two sets of images are
cropped and center projection transformed separately to obtain the region of interest. On this basis, the improved residual network model is used to construct different depot detection models for the two sets of images, respectively, and the above algorithm is verified by actual measurements.

Fig. 2 Inventory location with or without goods and surrounding inventory location (The five red identification areas are no goods, no goods, no goods, with goods, no goods)

Fig. 3 Inventory status and out of stock status (a and b are ground level, c and d are shelf level)
Clipping and center projection transformation

Each time the camera takes a picture of the warehouse location, the angle is fixed, and the relative position of the crossbeam where the goods are placed in the picture is unchanged. Therefore, the image can be cropped, and the image near the target warehouse location can be taken as the region of interest, which can not only reduce the interference of other regions, but also improve the accuracy of recognition. As shown in Fig. 3, with the change of angle, the image after direct clipping is not very regular, and other locations will affect the judgment of the current location. Therefore, this paper first cuts the image according to the location of the storage location, and then carries out the center projection transformation, so that any quadrilateral cut in the previous step can be transformed into 224 pixel × 224 pixel image. At the same time, in order to ensure the clarity of the final input classifier image, it is necessary to perform quadratic interpolation operation.

As shown in Fig. 5, in order to simplify the problem, assuming that the origin Q (0,0) of the square region coincides with R (0,0), it can be proved that any quadrilateral can be projected to the rectangle, and the transformation formula is independent of the coordinates of point E.

The coordinate of the vector $\mathbf{r}_{11}$ in the coordinate system $O_{10}O_{10}$ is $(a_0, a_1)$

$$r_{11} = a_0r_{10} + a_1r_{01} \quad (1)$$

Any point in a rectangular area:

$$q = x_0q_{10} + x_1q_{01} \quad (2)$$

Need to find the coordinates of the corresponding vector $\mathbf{r}$ in the coordinate system or $10r_{01} (y_0, y_1)$
Points €, € and € are collinear, so the coordinates of point € can be expressed as:

\[ r = y_0r_0 + y_1r_1 \]  

(3)

where € is the multiplication coefficient, so:

\[ r_{00} = (0, 0, 0) = E + t_{00}(q_{00} - E) \]  

(5)

\[ r_{10} = (1, 0, 0) = E + t_{10}(q_{10} - E) \]  

(6)

\[ r_{01} = (0, 1, 0) = E + t_{01}(q_{01} - E) \]  

(7)

\[ r_{11} = (1, 1, 0) = E + t_{11}(q_{11} - E) \]  

(8)

The zeros of the rectangle and the quadrilateral coincide, so \( t_{00} = 1 \).

Suppose that there is a normal vector \( n \) in the rectangular plane, then the left side of the above Eq. (6), Eq. (7), and Eq. (8) is multiplied by \( N(n_0, n_1, n_2) \), at the same time.

\[ (1 - t_{10})N \cdot E = n_0, \ (1 - t_{01})N \cdot E = n_1, \ (1 - t_{11})N \cdot E = n_0 + n_1 \]  

(9)

Add the left part of the equal sign of the two equations and compare with the third equation, the right part of the equation is \( n_0 + n_1 \). Therefore:

\[ t_{11} = t_{10} + t_{01} - 1 \]  

(10)

Substituting Eq. (2) and Eq. (10) into Eq. (8),

\[ (1, 1, 0) = (1 - t_{11})E + t_{11}(a_0r_{10} + a_1r_{01}) \]  

(11)

Namely:

\[ (1, 1, 0) = (2 - t_{10} - t_{01})E + (t_{10} + t_{01} - 1)(a_0r_{10} + a_1r_{01}) \]  

(12)

Substituting Eq. (7) from Eq. (8), we have:

\[ (2 - t_{10} - t_{01})E + t_{10}r_{10} + t_{01}r_{01} = (1, 1, 0) \]  

(13)

Subtracting Eq. (13) and Eq. (12):

\[ a_0(t_{10} + t_{01} - 1) - t_{10}r_{10} + a_1(t_{10} + t_{01} - 1) - t_{01}r_{01} = 0 \]  

(14)

So:

\[ t_{10} = a_0(t_{10} + t_{01} - 1) \]

\[ t_{01} = a_1(t_{10} + t_{01} - 1) \]  

(15)

The resulting coefficients are respectively:

\[ t_{00} = 1, \ t_{10} = \frac{a_0}{a_0 + a_1 - 1}, \ t_{01} = \frac{a_1}{a_0 + a_1 - 1}, \ t_{11} = 1 \]  

(16)

Substituting Eq. (4), we can get that if any point \((x_0, x_1)\) in the square area is known, we can find out the corresponding point coordinate \((y_0, y_1)\) (in or \(r_{01}\) coordinate system) in the quadrilateral area according to Eq. (17):

\[ (y_0, y_1) = \frac{(a_0x_0, a_1x_1)}{(a_0 + a_1 - 1) + (1 - a_1)x_0 + (1 - a_0)x_1} \]  

(17)

Generally, the value of \( a \) is not an integer, as shown in Fig. 6. Suppose that the calculated point \( P \) falls between the four adjacent pixels \( Q_{11}, Q_{12}, Q_{21}, \) and \( Q_{22} \). The simplest method is the nearest-neighbor interpolation method, that is, in the four neighboring pixels of the pixel to be solved, the gray level of the pixel nearest to the pixel to be solved is assigned to the pixel to be solved. Although the calculation amount of this method is small, it may cause discontinuity in the gray level of the image generated by interpolation, and obvious serration may appear where the gray level changes, as shown in Fig. 7. The calculation of bilinear interpolation method is more complex than that of nearest-neighbor point method, but the result is basically satisfactory. Figure 7 is a comparison picture of using nearest-neighbor interpolation and quadratic interpolation. It can be seen that using quadratic interpolation can improve the clarity of the picture.
According to the principle of quadratic interpolation, the pixel value of point $P$ can be calculated by the following formula:

$$P_P = w_0 h_0 P_{Q11} + (1 - w_0) h_0 P_{Q12} + w_0 (1 - h_0) P_{Q12} + (1 - w_0) (1 - h_0) P_{Q22}$$

where, $PQ_{11}$, $PQ_{12}$, $PQ_{21}$, and $PQ_{22}$ are the pixel values of pixel $Q11$, $Q12$, $Q21$, and $Q22$.

**ResNet-50**

Convolution neural network will degenerate when the number of layers becomes more; that is, when the number of layers reaches a certain number, the recognition accuracy will reach saturation. For example, on CIFAR-10, a small data set, the performance of 56-layer neural network is less than that of 20-layer neural network. It can be seen that the deeper the layers, the worse the network.

In order to solve the problem that the gradient disappears with the increase of layers, He Kaiming and others proposed ResNet to solve it. The basic idea is shown in the figure.

1. Fig. 8a: when adding a network, $x$ is mapped to output of $y = F(x)$.
2. Fig. 8b: improved figure a, output $y = F(x) + x$. In this case, the representation of output features is not learned $y$ directly, but learned $y - x$.

If you want to learn the representation of the original model, you only need to set all the parameters of $F(x)$ to 0, then $y = x$ is an identity mapping.

$F(x) = y - x$ is also called residual term. If the mapping of $x \rightarrow y$ is close to identity mapping, it is easier to learn residual term in Fig. 8b than complete mapping in Fig. 8a.

The structure in Fig. 8b is the basis of residual network. This structure is also called residual block. Input $x$ can propagate data forward or backward more quickly through cross layer connection. The specific design scheme of residual block is shown in the figure, which is also called bottleneck structure.

When downsampling is needed, the downsampling residual model can be used. The downsampling residual module is shown in Fig. 9. When downsampling is needed, for stacked convolution layers, downsampling can be achieved by setting the step size of the first convolution layer to 2. In order to prevent accuracy reduction, the number of filters is usually doubled, as shown in Fig. 9. For quick connection, by using a filter with size of $1 \times 1$ and step size of 2 (denoted as $1 \times 1@1/2$) By convolution,
the width and height of the input tensor are reduced to half of the original, and the depth is changed to 1. At this time, the shape of the output tensor of the shortcut link is not consistent with the shape of the output tensor of the convolution channel, and the tensor depth needs to be increased by filling 0. In other words, the network layers are added one by one in depth, so that the width and height of the added network layers are half of the original input tensor, and the elements are all 0, until the depth of the output tensor of the quick link is consistent with the convolution tensor.

At this time, the shape of the output tensor of the fast channel and convolution channel is exactly the same. First, add layer by layer and element by element, and then use the ReLU activation function to activate; the calculation output of the whole residual module is completed.

As shown in Fig. 10, it shows the structure of ResNet-50, which includes 49 layers of convolution and 1 layer of full connection, so it is called ResNet-50.

As shown in Fig. 11, this paper intercepts an empty location 224 × 224 × 3 of the pictures after the ResNet-50 layers of the picture.

Input the cut 224 × 224 image into the network. When the accuracy of the model is high and does not change in 5 h, the training will be stopped and the checkpoint file of the model will be exported, and then the checkpoint file will be converted into onnx file, which is convenient to call on the UWP platform.

Using the batch training model, the final training set is not obtained at the beginning. As shown in Fig. 12, the training steps are as follows:

1. Install the bracket on the roof of the warehouse and determine the installation position of the camera;
2. Collect, deduplicate, cut, and transform the warehouse pictures;
3. The ground layer and shelf layer images are separated to get the initial training set.
4. Recollection, deduplication, cropping, transformation, get a new group of pictures. The trained model is used to recognize these images.
5. If the probability of recognition result is less than 90%, the recognized image is added to the incremental training of the model.
6. Judge whether the number of pictures in the training set is enough. If the number of pictures is not enough, repeat steps (4) and (5). If the number of pictures is large enough, stop the training.

Experiment result

Data preparation

Because the warehouse needs to identify different areas, the shelf color cannot be unified. The goods are mainly composed of plastic buckets for storing liquids and cartons for storing solids and spade boards. According to the needs, the storage size will change frequently. Some storage tools, such as small forklifts, will also be stored in the cargo space. The camera is located directly above the roadway, and the camera fixing bracket is located on the ceiling. Because of the structure of the ceiling, the relative position between the camera and the shelf cannot be guaranteed to be the same. The north side of the warehouse is
a large outdoor unloading area. With the change of time, the light inside the warehouse changes all the time. Especially for the camera near the north side, the photo light of the warehouse near the north shelf is the strongest. At the same time, the situation of goods and no goods around a warehouse location also has a certain impact on the warehouse location. For example, if there are no goods in the upper warehouse location on the ground floor, there will be no shadow on the ground, and there will be a shadow when there are goods.

As shown in Table 1, about 2838 ground level images and 10,655 non-ground level images were collected initially. The number of images after multiple incremental training was 27,903 ground level images and 102,614 non-ground level images. According to the ratio of nearly 4:1, it is divided into train set and test set. The specific data are shown in Tables 2.

**Table 1. Initial picture allocation**

|                | Initial number of images | Train/test number of images |
|----------------|--------------------------|-----------------------------|
| Floor-layer    | 2838                     | 2255/583                    |
| Upper-layer    | 10,655                   | 8554/2101                   |
| Mix            | 13,493                   | 10,809/2684                 |

**Table 2. Picture distribution after expansion**

|                | Final number of images  | Train/test number of images |
|----------------|-------------------------|-----------------------------|
| Floor-layer    | 27,903                  | 22,317/5586                 |
| Upper-layer    | 102,614                 | 81,939/20,675               |
| Mix            | 130,517                 | 104,256/26,261              |
Experimental results and analysis

As shown in Fig. 13, the training test results of the ground layer, shelf layer and mixed layer are compared, respectively. The blue curve is the ground layer, the red curve is the shelf layer, and the black curve is the mixed picture. The dotted line represents the test result of more pictures. It can be seen that the accuracy of training according to the ground level and shelf level separately (96.055% and 95.859%) is better than that of training after mixing (94.821%). With the increase of images, the accuracy of ground level and shelf level training is improved, and the accuracy is close (99.119% and 99.352%), the difference is that the ground level training images are less than the non-ground level training images. After mixing, the accuracy of training decreased slightly (from 94.821% to 93.554%), indicating that there is interference between the ground and non-ground images, resulting in the decline of accuracy. Table 3 shows the specific test pictures and the corresponding accuracy.

We use the trained model in the actual production of the warehouse. The following table reflects the test results from September 15, 2020 to October 15, 2020. The cycle of the warehouse location identification system to collect the warehouse location pictures is about 9–10 min; 3246 photos are collected at a time, and the system collects them 24 h a day. As shown in Table 4, The recognition accuracy reaches 99.62%, which can fully meet the needs of daily storage production.

Network comparison

The accuracy comes from two aspects: specificity and sensitivity.

$$R_{TP} = \frac{TP}{TP + FN}$$

$$R_{TN} = \frac{TN}{TN + FP}$$

TP, FN, TN, and FP represent true positive, false negative, true negative, and false positive. RTP and RTN correspond to true positive rate (sensitivity) and true negative rate (specificity). Therefore, the accuracy can be calculated:

$$R_{AC} = \frac{TN + TP}{TN + FP + TP + FN}$$

As shown in Table 5, the sensitivity, specificity, and accuracy of AlexNet (VGG-19) (Krizhevsky et al. 2012), pointwise convolution (Xception) (CHOET 2016), dense connection (DenseNet) (Huang et al. 2016), and residual learning (ResNet-50) (He et al. 2016) were compared.

Table 3 Stratified and mixed test images and corresponding accuracy

| Layer       | Number of images | Final accuracy |
|-------------|------------------|---------------|
| Floor layer | 2838             | 96.055%       |
| Floor layer | 27,903           | 99.119%       |
| Shelf layer | 10,655           | 95.859%       |
| Shelf layer | 102,614          | 99.352%       |
| Mixture     | 13,493           | 94.821%       |
| Mixture     | 130,517          | 93.554%       |

Table 4 Identification result statistics of warehouse location identification system

| Location status | Number of samples | Identify the exact number | Recognition accuracy |
|-----------------|-------------------|---------------------------|----------------------|
| In stock        | 11,331,600        | 11,296,472                | 99.67%               |
| Out of stock    | 4,628,400         | 4,602,943                 | 99.45%               |
| Total           | 15,960,000        | 15,899,415                | 99.62%               |

Table 5 Experimental results of network identification

| Models     | Sensitivity | Specificity | Accuracy |
|------------|-------------|-------------|----------|
| VGG-19     | 87.32%      | 97.36%      | 88.19%   |
| ResNet-50  | 93.67%      | 98.20%      | 99.53%   |
| Xception   | 93.79%      | 96.69%      | 98.89%   |
| DenseNet   | 91.87%      | 97.10%      | 99.07%   |
The results show that ResNet-50 is the best in three aspects, and DenseNet, which is similar to DenseNet, has the same accuracy, but it needs higher computer performance, so this paper selects ResNet-50 as the model training method.

As shown in Fig. 14, ResNet-50 is compared with other residual networks. The number of pictures is 100, 1000, 10,000 and 100,000 at a time. It can be seen that the four test results of residual network of 34 layers are less than those of other layers. The test results of 101-layer residual network and 152-layer residual network are very close to those of 50-layer residual network.

Conclusions

In order to realize the intellectualization and unmanned of warehouse management system, this paper puts forward a kind of storage location state detection method based on ResNet-50. Firstly, the image acquisition of a multi-level shelf containing multiple bays is performed under different time and lighting conditions. Due to the difference in image characteristics between the bottom shelf on the ground and the upper shelf on the non-ground level, the collected images were divided into two groups: floor images and shelf images; and the warehouse status recognition was performed on the two groups separately. The two sets of images are cropped and center projection transformed separately to obtain the region of interest. On this basis, the improved residual network model is used to construct different depot detection models for the two sets of images, respectively, and the above algorithm is verified by actual measurements. We can draw the following four conclusions.

1. In terms of hardware layout, this paper selects the image signal with low cost and strong stability as the follow-up input. Using the Internet ball machine to get the image of each location from the top of the warehouse can not only clearly observe the location, but also not affect the loading and unloading operations.

2. In order to enhance the anti-interference ability of the detection method, this paper selects ResNet-50 network model to identify the inventory status. In order to enhance the stability of the model, on the one hand, the image is clipped and rotated during training and testing, so as to avoid inputting too much irrelevant information into the model; On the other hand, according to the image features of the location, the location is divided into bottom location and non-bottom location, and the model is trained, respectively. The experimental results show that the accuracy of clipping is 10% higher than that of clipping, and the detection result of location classification is 5% higher than that of non-classification.

3. Compared with other deep learning algorithms, the detection accuracy of this algorithm is the highest, which is 99.62%. In addition, the detection time is 20 ms, which can meet the actual needs of the site.

4. Through the follow-up investigation, this method improves the accuracy of warehousing, reduces the invalid and repeated running of forklift and other warehousing and handling tools, and indirectly reduces the use of lighting, air conditioning, display screen, and energy consumption in the warehouse.

This paper can identify the status of the location efficiently and accurately. Next, we are going to further study the location identification and classification from the perspective of this paper, which will provide more flexible support for the intelligent warehousing system.
Acknowledgements We are thankful to the anonymous reviewers for their detailed comments, although any errors are our own and should not spoil the reputations of these esteemed persons.

Author contribution MZL: conceptualization, investigation, methodology. XX: data collection, writing original draft, and review. XQW: conceptualization, resources, writing original draft, and review. CHL: methodology, resources, formal analysis.

All the authors read and approved the final manuscript.

Data availability The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate The authors whose names are listed immediately below certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

Author names:
Dr. Conghu Liu.
Dr. Qiannan Jiang.
Dr. Xiaoqiao Wang.
Dr. Xin Xu.
Dr. Mingzhou Liu.

Consent to participate Not applicable.

Consent for publication Not applicable.

Competing interests The authors declare no competing interests.

References

Alexander D, Timo S et al (2017) Recognizing grabbing actions from inertial and video sensor data in a warehouse scenario. Procedia Computer Science 110:16–23

An H, Razaq A, Nawaz A et al (2021) Nexus between green logistic operations and triple bottom line: evidence from infrastructure-led Chinese outward foreign direct investment in Belt and Road host countries. Environmental Science and Pollution Research. 1–24

Asaoka T, Nagata K, Nishi T et al (2018) Detection of object arrangement patterns using images for robot picking. ROBOMECH J 5(1)

Assia A, Noureddine EB, Abdelmoumen T et al (2018) An embedded system based on DSP platform and PCA-SVM algorithms for rapid beef meat freshness prediction and identification. Comput Electron Agric 152:385–392

Baesso RM, Oliveira P (2018) Using ultrasonic velocity for monitoring and analysing biodiesel production. Fuel 226:389–399

Chatpreecha P, Keatmanee C (2018) Stock monitoring unit in storage areas enable flexibility, productivity, and reliability of warehousing system

Chen Ch, Lei Zh (2019). Monitoring of contact state of GIS switch based on infrared sensing technology. J Eng

Chen X, Huang W (2020). Texture features and unsupervised learning-incorporated rain-contaminated region identification from x-band marine radar images. Mar Technol Soc J

Chollet F. Xception: Deep learning with depthwise separable convolutions. arXiv e-prints, 2016.

Dai Y, Gao HO (2016) Energy consumption in China’s logistics industry: a decomposition analysis using the LMDI approach. Transp Res Part D Transp Environ 46:69–80

Ding K, Jiang P, Su S (2018) RFID-enabled social manufacturing system for inter-enterprise monitoring and dispatching of integrated production and transportation tasks. Robotics & Computer Integrated Manufacturing 49:120–133

Fan W, Xu M, Dong X, Wei H (2017) Considerable environmental impact of the rapid development of China’s express delivery industry. Resour Conserv Recycl 126:174–176

Fergani C, Idrissi El Bouzekri El, A, Marcotte S et al (2020) Optimization of hyperconnected mobile modular production toward environmental and economic sustainability. Environ Sci Pollut Res 27:39241–39252

Gebhardt C, Trimborn T, Weber F et al (2020) Simplified ResNet approach for data driven prediction of microstructure-fatigue relationship

Gu X, Qu C (2020) A study of community surveillance system improvement based on ResNet person re-identification. Journal of Physics: Conference Series, 1575(1):012231 (6pp).

He K, Zhang X, Ren S et al (2016) Residual learning for image recognition [C]// IEEE Conference on Computer Vision & Pattern Recognition. IEEE Computer Society.

Hedjam R, Shaikh AK, Luo Z (2021) Ensemble clustering using extended fuzzy k-means for cancer data analysis. Expert Syst Appl 172(1):114622

Huang C, Qi J, Liu F (2015) Modeling emergence of network radar countermeasure system. Journal of System Simulation 27(6):1357–1367

Huang, Ling, J (2021) Measuring embodied carbon dioxide of the logistics industry in China: based on industry stripping method and input-output model. Environ Sci Pollut Res 28:52780–52797

Huang G, Liu Z, Laurens V et al (2016) Densely connected convolutional networks. IEEE Computer Society

Egger J, Michlmayr S (2018) A Faraday effect magnetic stripe scanner. J Phys: Conf Ser 1065:032013

Krizhevsky A, Sutskever I, Hinton G (2012) ImageNet classification with deep convolutional neural networks [C]// NIPS. Curran Associates Inc.

Lang Xu (2018) Decision and coordination in the dual-channel supply chain considering cap-and-trade regulation. J Clean Prod 197:551–561

Lang Xu et al (2021) Estimating the effect of COVID-19 epidemic on shipping trade: an empirical analysis using panel data. Mar Policy 133:104768

Lecun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7553):436

Liu H, Yao Z, Zeng L et al (2019) An RFID and sensor technology-based warehouse center: assessment of new model on a supermarket in China. Assem Autom 39(1):86–100

Liu, Alex X et al (2015) Fast and accurate estimation of RFID tags. IEEE/ACM Transactions on Networking: A Joint Publication of the IEEE Communications Society, the IEEE Computer Society, and the ACM with Its Special Interest Group on Data Communication 23(1):241–254

Liu C, Gao M, Zhu G, Zhang C, Zhang P, Chen J, Cai W (2021) Data driven eco-efficiency evaluation and optimization in industrial production. Energy 224:120170

Liukkonen M, Tsai TN (2016) Toward decentralized intelligence in manufacturing: recent trends in automatic identification of things. Int J Adv Manuf Technol 87(9/12):2509–2531
Mohamed L, Gunasekaran M et al (2020) Fighting against COVID-19: a novel deep learning model based on YOLO-v2 with ResNet-50 for medical face mask detection. Sustainable Cities and Society

Pablo F, Nedz (2016) SmartPort: a platform for sensor data monitoring in a seaport based on FIWARE. Sensors, 16(3):417

Pan TS, Huang H C, Lee J C, et al. (2020). Multi-scale ResNet for real-time underwater object detection. Signal Image and Video Processing, 1–9.

Rennie C, Shome R, Bekris KE et al (2016) A dataset for improved RGBD-based object detection and pose estimation for warehouse pick-and-place. IEEE Robotics & Automation Letters 1(2):1179–1185

Sangrok H, Yong Tae P (2019) Extending bluetooth LE protocol for mutual discovery in massive and dynamic encounters. IEEE Trans Mob Comput 18(10):2344–2357

Nathan S et al (2018) Expectation maximization-based satellite image segmentation. Journal of Engineering and Applied Sciences 13(12):9343–9345

Stephen K, Lyu D, Kenji T (2021) Vehicle detection and type classification based on CNN-SVM. International Journal of Machine Learning and Computing 11(4):304–310

Sun, X., Yu, H., Solvang, W.D. et al. (2021). The application of Industry 4.0 technologies in sustainable logistics: a systematic literature review (2012–2020) to explore future research opportunities. Environ Sci Pollut Res.

Tang Z, Bo L, Liu X, et al. (2021). An autoencoder with adaptive transfer learning for intelligent fault diagnosis of rotating machinery. Measurement Science and Technology, 32(5)

Tao Y (2021). Life as a self-referential deep learning system: a quantum-like Boltzmann machine model. Biosystems.

Tavana M, Tohidi H, Alimohammadi M et al (2021) A location-inventory-routing model for green supply chains with low-carbon emissions under uncertainty. In Press, Environmental Science and Pollution Research

Tianjian L, Bin H, at al. (2019) Application of convolution neural network object detection algorithm in logistics warehouse. The Journal of Engineering 23:9053–9058

Xu, Lang, F. Xie, and C. Wang. (2021). Passive or proactive capacity sharing? A perspective of cooperation and competition between two regional ports.

Xie X, Shao S, Lin B (2016). Exploring the driving forces and mitigation pathways of CO2 emissions in China’s petroleum refining and coking industry: 1995–2031. Appl Energy

Yesenia A, Michael A (2019) An automated supermarket checkout system utilizing a SCARA robot: preliminary prototype development. Procedia Manufacturing 38:1558–1565

Zhao K, Zhu M, Xiao B, et al. (2020). Joint RFID and UWB technologies in intelligent warehousing management system. IEEE Internet of Things Journal, PP (99):1–1.

Zheng H, Liu J, Ren X (2022) Dim target detection method based on deep learning in complex traffic environment. Journal of Grid Computing 20:1

Zhou L, Zhang Q (2021). Recognition of false comments in e-commerce based on deep learning confidence network algorithm. Information Systems and e-Business Management, 1–18.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.