Cargo pallets real-time 3D positioning method based on computer vision

Tianjian Li1, Qiu Jin1, Bin Huang2, Chang Li1, Min Huang1

1School of Mechanical and Electrical Engineering, Beijing Information Science and Technology University, Beijing 100192, People's Republic of China
2School of Automation Science and Electrical Engineering, Beihang University, Beijing 100191, People's Republic of China
E-mail: philip6907@163.com

Abstract: In storage environment, aiming at the problem of goods positioning when picking, the pallet is firstly recognised based on deep learning. Then, algorithm of obtaining the pose of the pallet by the image processing and Kinect sensor is proposed in this study. The pallet is recognised and its selected box is obtained by deep learning. On this basis, the position and the angle of the pallet are obtained by the image processing method, and then RGB-D transforms the position and posture of the pallet into the three-dimensional (3D) coordinate for three-dimensional positioning. The experiment results show that the algorithm can obtain real-time pallet position with the success rate of 81.02%. Thus, the algorithm can meet the requirements of the efficiency and accuracy location requirements of the storage of goods when picking.

1 Introduction
With the development of economic and scientific technology, logistics has entered a very rapid development and phase. Storage is a key part of that, and it is considered to be a third party profit for the corporation. Since it reduces the overall supply chain costs and improves the efficiency of the goods operation and the integrated service level [1]. At the present stage, one of the main problems in logistics is the identification and sorting of goods, such as warehousing and sorting, and traditional manual work. In view of the above problems, we identify the location of the goods by visual positioning, and then the warehouse robot will carry out the relevant sorting operation.

In recent years, RGB-D sensor has been widely used in the field of real-time positioning and three-dimensional (3D) reconstruction. RGB-D sensor can obtain the depth value of the scene by using depth ranging technology, thus obtaining the 3D information of the scene. The advantage of such sensors is to obtain 3D information in real time, which is convenient to carry and low cost. The typical RGB – D sensor has a Kinect and StructureSensor, and the RGB camera that has been used in this paper is a Microsoft Kinect, with an average accuracy of one millimetre.

In this paper, we propose to identify and identify pallet boxes by deep learning [2], and then further precise positioning by image processing, and obtain the space position of pallet cargo through 3D reconstruction.

2 Identification of target objects based on deep learning

2.1 Identification network of pallet cargo based on deep learning

A convolutional neural network is a kind of basic structure that applies a convolutional operation to a special context of a neural network to detect a convolutional neural network. The basic structure of a neural network is improved by referring to the SSD [3] detection algorithm. The first part of the network model is the basic network VGG16 [4], but the whole connection layer (such as FC6 and FC7 layer) used for classification after VGG16 is removed. After VGG16, the convolution layer is followed by Conv6, Conv7, Conv8_2 (including 1*1*256 and 3*3*512-s2) and Conv9_2 (including 1*1*128 and 3*3*256-s2). Its basic structure is as follows (Fig. 1).

2.2 Pallet identification

Input to any size image has depth of the trained neural network, by the convolution of the neural network after feature extraction, after according to VGG16, Conv7, Conv8_2 and Conv9_2 convolution layer to get the characteristics of the four different scale figure to predict the position of the object Bounding Box, and the corresponding confidence level, the final will show pallet's Bounding Box. Test end is shown in Fig. 2.

3 Pallet positioning method

3.1 Basic process of pallet positioning

The basic process of pallet positioning is shown in Fig. 3.

3.2 Edge detection

Edge detection is a common method of segmentation based on greyscale mutation. The pallet has obvious edge features in transverse and vertical directions, which is an important feature of pallets. In this paper, Canny edge detection is used to extract the edge features of pallets. Canny edge detection has the characteristics of low error rate, high positioning and minimum
response, and is considered to be the best edge detector. The calculation formula of Canny operator [5] is as follows:

\[
G_n = n \nabla G
\]

\[
\begin{align*}
\theta &= \arctan\left(\frac{\partial G}{\partial x}\right) \\
G &= \frac{\nabla G \ast I(x, y)}{\nabla G \ast I(x, y)}
\end{align*}
\]

where \( n \) is the edge direction, \( V_G \) is the vector for calculating the gradient, and the strength of the edge is determined by the item in formula \( \nabla G \ast I(x, y) \).

In this paper, the image is processed by OpenCV library, and the image after the Canny edge treatment is shown in Fig. 4.

### 3.3 Text-based image segmentation [6]

To further precisely determine the centre of the pallet, through analysing the pallet, the common pallet can be divided into fluctuation two parts, one part is on the horizontal straight lines, the following part is composed of vertical line, and the vertical lines is total of six. In order to obtain the position of the horizontal direction of the pallet, the pallet can be divided according to the shape of the pallet and the local characteristics. From the perspective of the texture characteristics of the pallet for the lower part of the pallet has a certain regularity, can use the statistical method of statistics from the ‘black’ to ‘white’ or ‘white’ to ‘black’ jump times [7], and set a threshold segmentation, pallet for setting the threshold value by analysing the pallet to 12, the specific algorithm process is as follows:

1. The binary image after input edge detection;
2. Scan the binary images in sequence from top to bottom, left to right. Count the number of jumps for each row \( M(i) \); The calculation formula of the jump of the \( I \) line is

\[
M(i) = \sum_{j=0}^{n} |f(j, i) - f(j + 1, i)|/255
\]

where \( w \) is the width of the image, \( J \) is the \( j \)th column of the image, \( f(j, i) \) is the pixel value of the image in \( (j, i) \);
3. From top to bottom, the total number of jumps in every 20 rows is counted as \( s(i) = M(i) + \ldots + M(i + 20) \), and find out the maximum number of jumps. Then record the boundary up = \( I \) and lower boundary drop = \( I + 20 \);
4. The upper boundary and the lower boundary are searched, respectively, and the upper boundary is moved up if the jumping variable of the upper boundary and adjacent rows is >12. Similarly, the lower boundary is shifted down. By searching the upper and lower boundaries, the upper boundary \( Y \) of the pallet is finally obtained.

The results of the algorithm are as shown in Figs. 5 and 6.

### 3.4 Inclination of pallet

The image in the upper part of the pallet image can be used to detect the inclination \( \theta \) of the pallet. The method of Hough transformation is used in image processing to detect a straight line, and then the inclination of straight line is determined. In this paper, the method of probability Hough transform [8, 9] is used to detect the tilt inclination \( \theta \) of the pallet. The Hough transform is a method used to find a straight line. Hough transformation uses the transformation between two coordinate spaces to map a curve or line with the same shape in one space to a point in another coordinate space to form a peak value, thus transforming the problem of detecting any shape into a problem of statistical peak value.

Through the accumulative chance of the Hough transform, it can detect a straight line on the edge of the pallet. Then using an operation, the HoughLinesP function, we can get two points \( (x_1, y_1) \), and \( (x_2, y_2) \) to calculate the inclination \( \theta \). The formula is as follows:

\[
\theta = \arctan\left(\frac{y_2 - y_1}{x_2 - x_1}\right)
\]

The results of the cumulative probabilistic Hough transform detection are as shown in Fig. 7.

### 3.5 Pallet centre coordinate

#### 3.5.1 Corrosion and expansion: The image in the lower part of the pallet image after segmentation can be used to detect the centre point of the pallet. In this paper, the accumulative probabilistic Hough transform is used to detect the lower part of the pallet. Through analysis, it is found that there is no complete connection in the middle of the edge line after Canny processing, and there is intermittent phenomenon, which will bring difficulties to the detection of the accumulative probabilistic Hough transform. Affect the accuracy of the test results. In this paper, the method of morphology is used to connect the edge of fracture, and it can reduce the interference in the linear detection process and improve the accuracy of the test results.
In this paper, by adopting the combination of inflation and corrosion [10] way to connect broken line, expands the structure of the element is $1 \times 3$, corrosion when the structure of the elements of $5 \times 1$. Dilate and erode are using in opencv function implementation. The results are shown in Fig. 8.

3.5.2 Find the midpoint of the pallet: After that, it is necessary to further extract the vertical straight line of the pallet and find two vertical straight lines in the middle, calculating the centre position of the pallet based on these two straight lines. Since each edge line could detect the straight line $>1$. So when looking for a centre of the pallet, must find a true straight line represents the edge of a pallet, then in incredible line sorting. Based on the above ideas, this paper proposes the following algorithm, whose basic process is as follows:

i. Use the HoughLinesP function to detect the edge of the lower part of the pallet, $(x_l, y_l)$ and $(x_r, y_r)$ are stored in the vector  $<CV::Vec4f> $ container;

ii. Sorting the detected lines by insertion sort method, and storing the sorted lines in $L$;

iii. Compare the distance between the lines in $L$, and if $<10$, the two lines $I$ and $J$ are the same line. Merge the two lines.

iv. Determine the number of lines in $L$, and if the number is odd, the middle line will be the centre line of the pallet, and the $Y$ will be brought into the line to obtain $X$; If the middle line of the two lines is calculated for even numbers, the line is the centre line of the pallet, and the value of $X$ can be obtained.

According to the above algorithm, the results are as shown in Fig. 9.

4 Pallet 3D positioning

Kinect [11, 12] its hardware structure is mainly divided into three parts, namely, a system-level chip, three cameras, and a four-element microphone array. Three of the cameras are colour cameras, infrared emitters, and infrared CMOS cameras. Through the infrared emitter signal, the infrared CMOS camera receives the infrared image.

With open source database OpenNI [13], the x-coordinate and y-axis coordinates of the depth image coordinate system can be converted to the $X$-axis and $Y$-axis coordinates in the world coordinates of the reference frame of the Kinect camera. The corresponding relationship between the camera coordinate system and the depth image coordinate system is shown in Fig. 10.

It is apparent to see that the coordinates of the world at some point are $(x_u, y_u, z_u)$, and the coordinates of the camera are $(x_c, y_c, z_c)$, and the coordinates of the image are $(x_v, y_v)$, and according to the traditional method of calibration [14], and the image coordinate system into one point of world coordinate conversion relations as follows:

$$
\begin{bmatrix}
    x_l \\
    y_l \\
    1
\end{bmatrix} = \begin{bmatrix}
    f_x & 0 & u_0 \\
    0 & f_y & v_0 \\
    0 & 0 & 1
\end{bmatrix} \begin{bmatrix}
    x_c \\
    y_c \\
    1
\end{bmatrix}
$$

where $f_x = f/p_x$, $f_y = f/p_y$, $f$ is the focal length of the camera, $p_x$ and $p_y$ represent the width and height of the unit pixels, respectively, $u_0$ and $v_0$ are the offset of the $X$-axis and $Y$-axis of the imaging centre point of the camera. $R$, $t$ represent the rotation and the translation vector of the camera relative to the world coordinate system, respectively.

It can be seen from Figs. 11 and 12 that that world coordinate system coincide with that of the camera, and the rotation matrix $R$ is the unit matrix, and the translation matrix is a zero matrix, which can be converted into the following relationship:

$$
\begin{bmatrix}
    x_u \\
    y_u \\
\end{bmatrix} = \begin{bmatrix}
    x_c - u_0 \\
    y_c - v_0
\end{bmatrix} \cdot \frac{1}{f_x}
$$

Since the camera and the colour RGB camera for the depth information shoot objects from a different inclination, the coordinates for the original data obtained are not yet calibrated. Through OpenNI’s correction function, the parallax problem caused by two cameras can be fixed, so that the target object can be aligned to the same position so that the subsequent mapping operation can be convenient.
The coordinate points from the centre of the pallet, which are available from Section 3.5 in this paper, are brought into the function [15] in the Kinect itself, with the following section codes:

```c
hResult=pCoordinateMapper->MapColorFrameToCameraSpace(depthWidth * depthHeight, &depthBuffer[0], colorWidth * colorHeight, cameraSpacePoints);
if (SUCCEEDED(hResult))
{
    long colorIndex = (long)((r.y) * colorWidth + r.x);
    CameraSpacePoint csp = cameraSpacePoints[colorIndex];
}
```

5 Experiment and results

5.1 Experimental environmental

In this experiment, visual studio 2013 was used as the experimental platform, and OpenCV2.4.9 version of the library was used as the tool for image processing. The experimental equipment mainly included the Kinect and the experiment pallet composition (Fig. 13).

5.2 Pallet positioning

When it can complete the observed pallet, placed to access the camera and the pallet in a fixed position, gathering every frame image, some experimental results as shown in Fig. 14, statistically its recognition success rate as shown in Table 1. According to the prior experience in the actual project, the middle point of recognition can be located in the middle of the pallet. In this experiment, the thickness of the pallet was 20 mm. Also the allowable error $\Delta = \pm 10$ mm was obtained by the experiment. The ideal location and experimental error range are as shown in Fig. 15.

In order to verify the accuracy of this experiment, the experimental data will be divided into two groups: one group will be used to verify the accuracy of pallet identification by changing the relative position of the Kinect camera and the pallet to the pallet inclination $\theta = 0$. Another set of the Kinect camera and the relative position of the pallet must be used to verify the accuracy of pallet identification by changing the size of the pallet inclination $\theta$. Considering the situation of the warehouse robot in the actual working process, it is affected by the actual storage environment and other factors. In the first experiment, the experiment was conducted at 1, 1.5, 2 and 2.5 m, respectively. For the second experiment, when the pallet was larger than 45°, the storage robot was very difficult and did not meet the actual working conditions. In the second experiment, the relative position of the Kinect camera and the pallet was set to 1.5 m, while the inclination of the pallet relative to the camera was set to 0°, ±30°, ±45°, looking down and looking up, respectively.

As shown in Fig. 16, some experimental results are obtained in different inclinations. Through the experiment, in the field of error, the method that we present is a very good way to adapt to the storage environment and to be able to do 81.02% of the time. At the same time, a good recognition effect can be obtained at different distances and the inclination of the pallet.

6 Conclusion

In this paper, the positioning problem of the storage robot is studied, and the method of image processing is proposed, and the 3D positioning algorithm is made through the Kinect sensor. It divides that image into two parts by edge detection and texture hop method, and the upper part is directly tested by the Hough transform inclination A, and after the morphological treatment of the following part, the central point of the pallet is found through the present algorithm. The Kinect sensor then converts 2D coordinates into 3D coordinates. By analysing the experimental data, the image recognition rate of this algorithm is 81.02%, which basically satisfies the positioning of the pallet. By analysing Tables 2 and 3, it is easy to see that the algorithm can accurately locate the pallets within the range of errors allowed by experiments.

---

Table 1  Success rate of image acquisition

| Captured images times | Successful identification times | Success rate, % |
|-----------------------|-------------------------------|----------------|
| 490                   | 397                           | 81.02          |

---

Fig. 12 Coordinate relationship between pallet image and Kinect

Fig. 13 Laboratory environment

Fig. 14 Actual laboratory environment

Fig. 15 Pallet error analysis

Fig. 16 Different inclinations are part of the experimental results
In the fourth step of finding the middle point of the pallet, as suggested in Section 3.5, if two or one straight line is not the straight line of the middle position of the pallet, then straight line in the middle of the straight line in $L$ can be used as the basis of the middle point of the pallet and the error can be introduced to the middle point of the pallet. In the following research, we will focus on how to accurately extract the middle line of the pallet to improve the accuracy of the detection.

7 Acknowledgments

Beijing Science and Technology Project (Z171100000817006). Beijing Municipal high level innovation team construction project (IDHT20180513).

8 References

[1] Xue, Y.J.: ‘Research on simulation and application of warehouse and logistics system’ (Dongnan University, China, 2006)
[2] Zheng, Y.F., Zhang, X.W., Cao, T.Y. et al.: ‘The application of deep convolutional neural network in image segmentation’, Military Commun. Technol., 2016, (2), pp. 98–104
[3] Liu, W., Anguelov, D., Erhan, D., et al.: ‘SSD: single shot multibox detector’. European Conf. Computer Vision, Amsterdam, Netherlands, 2016, pp. 21–37
[4] Simonyan, K., Zisserman, A.: ‘Very deep convolutional networks for large-scale image recognition’. arXiv preprint arXiv, 2014, pp. 1409–1556
[5] Xie, Z.M., Gao, X.D.: ‘The edge extraction technology of weld image based on Canny operator’, Weld. J., 2006, (1), pp. 29–32 + 36 + 114
[6] Kang, J.X.: ‘Design and realization of license plate recognition system based on image’ (Jilin University, China, 2014)
[7] Liu, G.Q., Zheng, X.H., Zhang, X.B.: ‘The license plate localization algorithm based on image texture feature extraction’, J. Chinese Image Graphics, 2005, (11), pp. 1419–1422
[8] Hu, B., Zhao, C.H.: ‘Fast lane line detection method based on probabilistic Hough transform’, Microelectron. Comput., 2011, (10), pp. 177–180
[9] Zhao, W.W.: ‘Based on the improved Hough transform detection method’, Manuf. Autom., 2010, (6), pp. 243–246
[10] Sui, H.Y.: ‘Investigation of erosion mechanism and protective technique for expansion joint of chemical pipelines’ (Daqing petroleum college, China, 2009)
[11] Ye, R.Z.: ‘Investigation of 3D reconstruction technique based on that depth sensor of the Kinect’ (Huananligong University, China, 2013)
[12] Fan, Z.: ‘3D reconstruction based on Kinect’ (Xidian University, China, 2013)
[13] Deng, R.: ‘The study and implementation of the gesture recognition system for the 3D scene video conference’ (Shanghai Jiaotong University, China, 2013)
[14] Zhang, Z.: ‘Flexible camera calibration by viewing a plane from unknown orientations’. Proc. 7th Int. Conf. Computer Vision, Greece, Corfu, 1999, pp. 666–673
[15] Xia, W.L., Gu, Z.P., Yang, T.S.H.: ‘The realization of real-time 3d reconstruction algorithm – based on the 3D reconstruction of the Kinect and the single vision SLAM’, Comput. Eng. Applic., 2014, (24), pp. 199–203

| Table 2  | Average absolute error of relative position between Kinect camera and pallet |
|---------|--------------------------------------------------|
| Relative position, m | Average absolute error, mm |
| 1      | 3.61 |
| 1.5    | 4.01 |
| 2      | 4.15 |
| 2.5    | 4.9  |

| Table 3  | Mean absolute error test of pallet |
|---------|----------------------------------|
| Pallet inclination $\theta$, deg. | Mean absolute error, mm |
| 0      | 3.88 |
| 30     | 7.61 |
| 45     | 9.40 |
| $-30$  | 7.23 |
| $-45$  | 9.37 |
| look down | 4.78 |
| look up | 4.13 |

This is an open access article published by the IET under the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0/)