An Object Recognition and Volume Calculation Method Based on Yolov3 and Depth Vision

Y Sun1, Z X Liu1, M Li1, Z T Zeng2, Z X Zong1, and C L Ji1,a

1 Shanghai Dian Ji University, Shanghai 201306, China
2 All Rank Vision Co., Ltd., Shanghai 201306, China
aCorresponding author’s e-mail: jicl@sdju.edu.cn

Abstract. In recent years, the logistics industry is in its full swing. The demand for logistics is increasing throughout the country, and this rising trend will continue during the following decades. With the substantial increase of workloads, traditional manual ways are increasingly unable to meet the new needs of the logistics industry, therefore new production and transportation modes are becoming imperative. Considering this, a new method was proposed to help detect, recognize and calculate the volume of rectangular objects automatically. In this method, a YOLO-based model was firstly applied to detect the object. Secondly, an RGB-D camera was used to obtain the depth information. Finally, after the length, width and height values were estimated from a single surface, the volume of the object can be calculated. Experimental results showed that the method is simpler and more accurate as compared with traditional three-dimensional reconstruction-based methods.

1. Introduction

With the popularity of e-commerce platforms and changes in people’s lifestyles, the logistics industry has also developed in full swing, and traditional manual operation methods can no longer meet the new needs of the logistics industry. In recent years, many express companies have started to build unmanned logistics sorting stations, in which small robots are used to identify parcels and classify them according to delivery address. However, very few research focuses on how to make the full usage of the delivery van’s load. Most companies arrange the loadings manually by means of contact-type volume calculation method. In order to overcome this, Yizhou Mi proposed an automatic volume calculation method. A semi-global dynamic programming algorithm was firstly applied to realize the stereo matching of parcel binocular images. Then the dense disparity map was obtained, which contains the three-dimensional (3D) spatial information of the parcel. Finally, real-time measurement of the parcel volume and 3D size was achieved through 3D reconstruction algorithm. This method works good for medium or large static objects, but the errors are quite big for small dynamic objects. Ding et al. extracted the 3D information of parcels by means of stereo matching and 3D reconstruction. By using a threshold segmentation algorithm, the average height and area of the parcel were obtained. Then the volume can be calculated through height and area values. The error rate of this algorithm is less than 5%. However, it only works fine for bulky objects like stacked materials. Although the volumes of express parcels variate frequently, they are small in most cases. So traditional methods are ineffective most of the time.

This paper proposed a parcel detection and volume measurement method based on You Only Look Once (YOLO) v3 network and depth vision. It can be divided into two parts. First thing is to establish
the object dataset and train the YOLO model. Then the model was used to detect cuboid parcels. After the approximate parcel area was detected, that is, a bounding-box of the image pixel range, the area will be processed later. Secondly, a depth camera was used to obtain the distance information of the corresponding pixels from RGB images after calibration. The length and width of the parcel can be calculated from the distance between the point on object surface and the camera. Experimental results showed that the proposed method can detect and measure volume of parcel object easily. When calculating the volumes of compact cuboid objects, the error rate is within 4%, which meets the practical demands of logistics industry.

2. Object detection

Object detection, segmentation and classification are three main tasks in the field of computer vision. As one of the most important branches of artificial intelligence (AI), machine vision has extensive applications in object recognition and classification areas which are widely used in all trades.

YOLO is a typical one-stage detection algorithm, which was originally proposed by Joseph Redmon in 2015 to achieve end-to-end target detection based on a single CNN model \[5,6\]. Although v4 and v5 have been released now, YOLOv3 is currently the most widely used one in engineering. The main performance indices for object detection model are detection accuracy and detection speed. For accuracy index, positioning accuracy of the object is of most importance, and classification accuracy comes next. Figure 1 compares the performances of typical object detection algorithms\[7\]. YOLOv3 not only has a higher accuracy in multi-category object detection, but also has a faster speed than other algorithms.

![Figure 1. Performance comparisons of typical object detection algorithms.](image)

3. Volume calculation

3.1 Edge extraction

It is assumed that the height between the fixed position of the camera and the test platform is \(h_c\). The resolution of the image taken by the camera is \(N \times M\). After the calculation of the approximate position of the object through the YOLO network, two points \((x_1, y_1)\) and \((x_2, y_2)\) in the pixel coordinates system can be obtained. Point \((x_1, y_1)\) is in the upper left corner, while point \((x_2, y_2)\) is in the lower right corner. In order to make sure the object to be detected is in the center of the image and to reduce the influences of the complex environment, the detection area (bounding box) given by YOLOv3 can be expanded as follows.

Firstly, the midpoint of the above two points can be obtained and named as \((x_{mid}, y_{mid})\). The length \(l\) and width \(w\) of the bounding box are defined in equation (1).

\[
\begin{align*}
    l &= x_2 - x_1 \\
    w &= y_2 - y_1
\end{align*}
\] (1)

By taking \((x_{mid}, y_{mid})\) as the center point and expanding the length and width of the bounding box to 1.2 times of their original sizes, the new upper left point \((x'_1, y'_1)\) and lower right point \((x'_2, y'_2)\) of the bounding box become \((x_{mid} - 0.6l, y_{mid} - 0.6w)\) and \((x_{mid} + 0.6l, y_{mid} + 0.6w)\) respectively. Here, 1.2 is an
empirical value. If any of the two vertices is beyond image boundary, its adjacent boundary pixel serves as a substitute. The calculation is shown in equation (2) and (3) respectively.

\[
\begin{align*}
    &x_i = 0, \text{ if } x_{mid} - 0.6l < 0 \\
    &y_i = 0, \text{ if } y_{mid} - 0.6w < 0 \\
    &x_i = M, \text{ if } x_{mid} + 0.6l > M \\
    &y_i = N, \text{ if } y_{mid} + 0.6w > N
\end{align*}
\]  

Then the new image is processed by the Canny operator to detect the area with most edge information. And the coordinates \((u,v)\) of edge points are recorded as a data set. The offsets \(x_{mid} - 0.6l\) and \(y_{mid} - 0.6w\) should be added to \(u\) and \(v\) respectively in order to align the bounding box coordinates to the original image coordinates (adding 0 instead when \(x_{mid} - 0.6l\) or \(y_{mid} - 0.6w\) is smaller than 0). After getting the depth image from the RGB-D camera, a new data set is defined as \(\{(u_i, v_i, d_i), (u_2, v_2, d_1), \ldots, (u_n, v_n, d_n)\}\).

3.2 Unit conversion

The values in logical unit differ from those in physical unit. Therefore, a relationship between them needs to be built before calculating the practical volume of the object. Firstly, by comparing the values of \((u', v')\) in pixel coordinates, the minimum and maximum value points and midpoint can be obtained as \((u'_\min, v'_\min)\), \((u'_\max, v'_\max)\) and \((u'_\text{mid}, v'_\text{mid})\) respectively. With vertical coordinate of the midpoint \(v'_\text{mid}\) getting fixed, its corresponding horizontal coordinates can be easily found in the frame data set, and the left and right points are recorded as \((u'_\text{mid}_\text{left}, v'_\text{mid})\) and \((u'_\text{mid}_\text{right}, v'_\text{mid})\) respectively. Furthermore, to reduce the impacts of noises and object shape, the middle point between \((u'_\text{mid}, v'_\text{mid})\) and \((u'_\text{mid}_\text{left}, v'_\text{mid})\) is denoted as \((u'_\text{mid}1, v'_\text{mid})\), and the middle point between \((u'_\text{mid}, v'_\text{mid})\) and \((u'_\text{mid}_\text{right}, v'_\text{mid})\) is denoted as \((u'_\text{mid}2, v'_\text{mid})\), as is shown in figure 2(a).

![Image](image.png)

(a) Image edge processing

(b) Unit conversion from the image

Figure 2. Image unit conversion.

Secondly, after obtaining the corresponding distance information according to the pixel coordinates, we denote them as \(d_1\), \(d_2\) and \(d_3\) respectively. In the same way, we denote the actual distance between two adjacent points as \(s\), and the angle between the camera and farthest point of the object plane as \(\theta\), as is shown in figure 2(b). According to the Law of Cosines, equation (4) can be defined as below:

\[
\begin{align*}
    \cos \theta &= \frac{d_1^2 + s^2 - d_3^2}{2d_1s} \\
    \cos \theta &= \frac{d_2^2 + 4s^2 - d_3^2}{4d_2s}
\end{align*}
\]
After combining two equations, we can obtain the relationship between $s$ and $d$ as equation (5):

$$s = \sqrt{\frac{d_1^2 - 2d_2^2 + d_3^2}{2}}$$

Then, the physical distance represented by a pixel is given by

$$\text{factor} = \frac{s}{u_{\text{mid}} - u_{\text{mid}}}$$

The value of the angle $\theta$ can be obtained by substituting the value of $s$ into equation (4).

### 3.3 Volume calculation

In order to calculate the object volume, we need to know the coordinates of the four vertices. Since the extreme values of $u'$ and $v'$ are $u'_{\text{min}}$, $v'_{\text{min}}$, $u'_{\text{max}}$ and $v'_{\text{max}}$, the coordinates of the four object vertices can be found in the edge data set and recorded as $(u'_{\text{min}}, v'_{\text{min}})$, $(u'_{\text{max}}, v'_{\text{max}})$, $(u'_{\text{mid}}, v'_{\text{min}})$, $(u'_{\text{mid}}, v'_{\text{max}})$.

The physical length and width of the object are given by:

$$
\begin{align*}
L & = \sqrt{(u'_{\text{max}} - u'_{\text{min}})^2 + (v'_{\text{mid}} - v'_{\text{mid}})^2} \times \text{factor} \\
W & = \sqrt{(u'_{\text{max}} - u'_{\text{mid}})^2 + (v'_{\text{mid}} - v'_{\text{min}})^2} \times \text{factor}
\end{align*}
$$

The vertical distance between the camera and the surface of the object is given by:

$$h_d = \sin \theta \ast d_i$$

Because the height of the camera is $h_c$, the height of the object is $h_c - h_d$, and the object volume is

$$V = (h_c - h_d) \ast L \ast W$$

### 4. System construction and experimental results

The experiment was conducted on a PC with an Intel Core I7 CPU and a NVIDIA RTX 2070 GPU. An Intel RealSense D435i camera was used to capture the RGB and depth images. The model training was completed on Ubuntu 16.04 OS, and the rest of the work was completed on Windows 10 OS.

#### 4.1 Object recognition results

Four different parcel boxes are used as the test objects. Training set is consisted of 200 images of four boxes which are labeled with "box1", "box2", "box3" and "box4" respectively. Figure 3(a) shows the default results given by YOLOv3, while figure 3(b) shows the improved results after expanding the bounding box. It is obvious that, after expanding all the objects are within the bounding box, which will help the edge detection later.

![Bounding box given by YOLOv3](image1)

(a) bounding box given by YOLOv3

![Bounding box after expanding](image2)

(b) bounding box after expanding

Figure 3. Boxes recognition results (arbitrary surface).
Image test set is consisted of 120 images. The four boxes in the test set are the same as those in the training set. There are 30 images for each box, with each surface captured 5 times in different postures. Table 1 shows the recognition accuracy and average maximum approximate probability (MAP) of the training model. The results of "box1" and "box2" are better because there are more image features on their surfaces.

Table 1. Comparison of boxes recognition results (arbitrary surface).

| Test image category | Quantity | Recognition accuracy (%) | Average MAP (%) |
|---------------------|----------|---------------------------|-----------------|
| box1                | 30       | 98                        | 80              |
| box2                | 30       | 90                        | 73              |
| box3                | 30       | 83                        | 51              |
| box4                | 30       | 86                        | 64              |

In this experiment, since the front surface of each box has a larger area and contains more textures as compared with other surfaces, the recognition accuracy and MAP are much better. Based on this, the following volume calculation will be performed in image front view.

Table 2. Comparison of boxes recognition results (in front view).

| Test image category | Quantity | Recognition accuracy (%) | Average MAP (%) |
|---------------------|----------|---------------------------|-----------------|
| box1                | 25       | 100                       | 99              |
| box2                | 25       | 96                        | 95              |
| box3                | 25       | 96                        | 93              |
| box4                | 25       | 96                        | 98              |

4.2 Volume calculation results

After the object detection, the length, width and height information of the object is obtained. The results and error analysis are shown in table 3 and figure 4 respectively.

Table 3. Volume calculation results (in front view).

| Box | Length (cm) | Width (cm) | Height (cm) | Volume (cm$^3$) |
|-----|-------------|------------|-------------|-----------------|
| 1   | 14.13       | 8.61       | 3.30        | 401.48          |
|     | Average test value | 14.20       | 8.65       | 3.37           | 413.94          |
|     | Error (%)   | 0.49       | 0.46        | 2.08           | 3.01            |
| 2   | 8.07        | 5.17       | 5.15        | 214.87          |
|     | Average test value | 8.02       | 5.20       | 5.21           | 217.28          |
|     | Error (%)   | 0.62       | 0.57        | 1.15           | 1.10            |
| 3   | 7.03        | 5.12       | 5.13        | 184.65          |
|     | Average test value | 7.00       | 5.10       | 5.10           | 182.07          |
|     | Error (%)   | 0.43       | 0.40        | 0.59           | 1.42            |
| 4   | 12.34       | 10.28      | 5.25        | 665.99          |
|     | Average test value | 12.40      | 10.33      | 5.32           | 681.45          |
|     | Error (%)   | 0.48       | 0.48        | 1.32           | 2.27            |
4.3 Result analysis
Experimental results show that the detection effect is good for objects with abundant features. Therefore, it is always better to measure boxes in front view. YOLOv3 works good for box detection and recognition. At the same time, according to table 3 the error of 3D sizes is at the level of millimeter and the volume error rate is below 4%. As is shown in figure 4, the error of this method meets the current industry's requirements for production and transportation.

5. Conclusion
This paper presents a cuboid object recognition and volume measurement method based on YOLOv3 and depth vision. As compared with traditional 3D reconstruction-based methods, its principle and calculation are much faster and simpler. The object can be located in the camera in real time, and the volume calculation error (4%) meets the practical demands. The proposed method works well as applied to measure compact cuboid objects. It can be used in the express distribution and loading industry.

References
[1] Yan Li. Current Situation and Development Trend of China's Logistics Industry [J]. Modern Business Trade Industry,2020,41(22):36-37.
[2] Weichang Zhang. Digital Innovation and Full Link Upgrade——Discussion on AI Empowered Logistics Transfer Center [J]. China Security & Protection,2020(07):70-73.
[3] Yizhou Mi. Express Package Volume Measurement System Based on Binocular Vision [D]. Hefei University of Technology,2017.
[4] Siyu Ding, Hongxia Miao, Bensheng Qi, Heng Zhang. Research on Volume Measurement of Irregular Stack Based on Binocular Vision [J]. Computer Measurement & Control,2020,28(04):71-74+84.
[5] Redmon J, Divvala S, Girshick R, etal. You only look once: Unified, real-time object detection. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA. 2016. 779–788.
[6] Redmon J, Farhadi A. Yolo9000: Better, faster, stronger. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA. 2017.6517–6525.
[7] Redmon J, Farhadi A. YOLOv3: An incremental improvement. arXiv preprint arXiv: 1804.02767, 2018.