Complex & Intelligent Systems (2021) 7:1759–1765
https://doi.org/10.1007/s40747-020-00182-z

ORIGINAL ARTICLE

Recognition and location of typical automotive parts based on the RGB-D camera

Wei Liu[1,2] · Fengpeng Li[1,2] · Cheng Jing[1] · Yidong Wan[1] · Beibei Su[1] · M. Helali[2]

Received: 17 March 2020 / Accepted: 17 July 2020 / Published online: 1 August 2020
© The Author(s) 2020

Abstract
Aiming at the problem that the accuracy of multi-part automatic assembly line sorting is not high, a set of machine vision-based recognition and positioning system is designed with KINECT as the RGB-D camera. The internal and external parameters of the RGB-D camera were calibrated using MATLAB; taking the automobile tire as the target part, because it is better for the system and more accurate, the feature invariant feature transformation (SIFT) algorithm is used to extract and match the feature points of the target part. The depth-based image obtains the spatial position parameters of the target part, thereby calculating the three-dimensional coordinates of the target part, and realizes the recognition and positioning functions of the system. The experimental results show that the visual positioning effectiveness is 96% in the unstructured indoor environment, and the system has good robustness and real-time performance.

Keywords RGB-D camera · Recognition and location · SIFT algorithm · Automatic assembly line

Introduction

Aiming at the low recognition consistency, sorting automation and intellectualization of multi-part automatic assembly line, machine vision is introduced to acquire color depth image (RGB-D) and analyze target features and depth, calculate the three-dimensional position of target parts, and realize the functions of automatic grasping and assembly of multi-part automatic assembly line.

Kumawat [1] proposed a fast feature point detection algorithm, and the results show that the time required to detect five feature points is relatively short in the case of mixed feature detector. Hu [2] proposed and evaluated several most advanced feature description algorithms, providing some guidance to design new feature description algorithms. Tang [3] proposed a robust matching method that filtered out the least important local features before matching, effectively ensuring the performance of local feature matching. Martin [4] constructed geometric features by discovering existing relationships between data. Shi [5] designed a unique matching algorithm based on the grid topology. The results show that the local features with multi-line descriptors are more robust than other classic features based on patch. Li [6] proposed a Harris multi-scale corner detection algorithm based on contour transformation. The detected corner points are more uniform and reasonable, which can be used in many fields such as image Mosaic. Liu [7] used continuous data frames to build sub-maps to get the general modeling method of parallel manipulators. Endres [8], using the distance between the feature descriptor to generate a depth image of the minimum spanning tree, removed the time close to the data and then randomly selected the method of random forest K frame data to detect loop-back, to some extent to meet the needs of real-time detection. Guo [9], using the algorithm of Tri SI (Tri-Spin-Image) with the combination of different weights, selected the best neighborhood structure and improved the local coordinate system. Salti [10] proposed SHOT features (Signature of Histograms of Orientation), where the feature histograms are obtained by constructing point cloud topological structure, with good rotation invariance and robustness.

In this paper, KINECT is taken as the research object [11, 12]. Firstly, the inside and outside parameters of the camera are calibrated by checkerboard calibration method; secondly, the feature points of the target parts are extracted

✉ Fengpeng Li
494533048@qq.com

[1] School of Automotive Engineering, Yancheng Institute of Technology, Yancheng 221051, Jiangsu, China

[2] Jiangsu Coastal Institute of New Energy Vehicle, Yancheng 221051, Jiangsu, China

by the SIFT algorithm, and the similarity between the feature points and the feature points of the target parts is compared to realize the recognition of the target parts; finally, the target parts are identified by depth image. By calculating the position parameters of the target parts, the three-dimensional coordinates are obtained to complete the positioning. Relevant research can improve the efficiency of multi-part automatic assembly line recognition and positioning, and provide theoretical basis and experimental support for the realization of intelligent automatic assembly line.

Calibration of internal and external parameters of the RGB-D camera

There are two main reasons for camera calibration: one is that the distortion degree of each lens is different in the process of production and assembly, which can be corrected by camera calibration to generate image-corrected lens distortion; the other is to build camera imaging geometric model according to the camera parameters obtained after calibration. The three-dimensional scene is reconstructed from the acquired image [13–15].

Camera model refers to the transformation relationship between the actual spatial points and the corresponding points of two-dimensional images. The internal and external parameters of RGB-D are obtained by using a simple and practical camera model and camera calibration method. In this paper, the pinhole model is selected to calculate the four coordinate systems and their transformation relations. The four reference coordinate systems are the world coordinate system, camera coordinate system, image coordinate system and pixel coordinate system.

Without considering the distortion, the three-dimensional space points are represented as P_u (X_u, Y_u) in the image coordinate system. Taking the upper left corner of the imaging plane as the origin of coordinate and the unit of pixels, the coordinates of the pixels are represented by (U, V) [16–18]. Both image coordinates and pixel coordinates are on the camera imaging plane, but the origin is different from the unit. Using d_x and d_y to represent the size of a pixel unit in the X and Y axis directions of the image plane, the transformation relationship between the image coordinates and the pixel coordinates can be obtained as follows:

$$\begin{cases} u = \frac{x}{d_x} + u_0 \\ v = \frac{y}{d_y} + v_0 \end{cases}. \tag{1}$$

The above formula is expressed by a matrix as follows:

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{d_x} & 0 & u_0 \\ 0 & \frac{1}{d_y} & v_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}. \tag{2}$$

According to the geometric relationship shown in Fig. 1, the transformation relationship between the image coordinate system and camera coordinate system can be obtained by the formula 3:

$$\begin{cases} x = \frac{f}{Z_c} X_c \\ y = \frac{f}{Z_c} Y_c \end{cases}. \tag{3}$$

The matrix is expressed as:

$$Z_c = \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \begin{bmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} X_c \\ Y_c \\ Z_c \\ 1 \end{bmatrix}, \tag{4}$$

where f denotes the focal length. The conversion between camera coordinate system and world coordinate system can be expressed by formula 5:

$$\begin{bmatrix} X_c \\ Y_c \\ Z_c \\ 1 \end{bmatrix} = \begin{bmatrix} R & t \\ 0^r & 1 \end{bmatrix} \begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix}. \tag{5}$$

Obtained from formulas 3, 4 and 5,

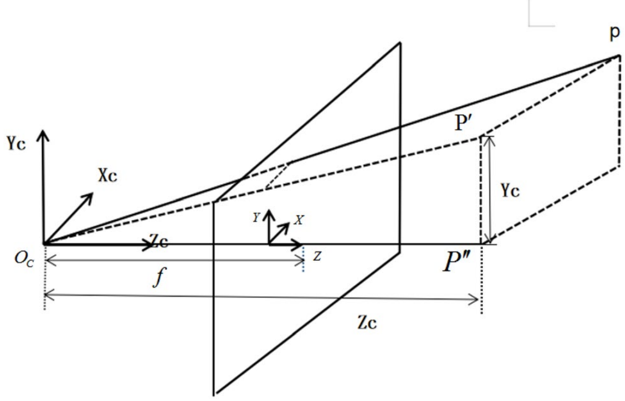


Fig. 1 Schematic diagram of the geometric relationship between the imaging plane and the camera coordinate system

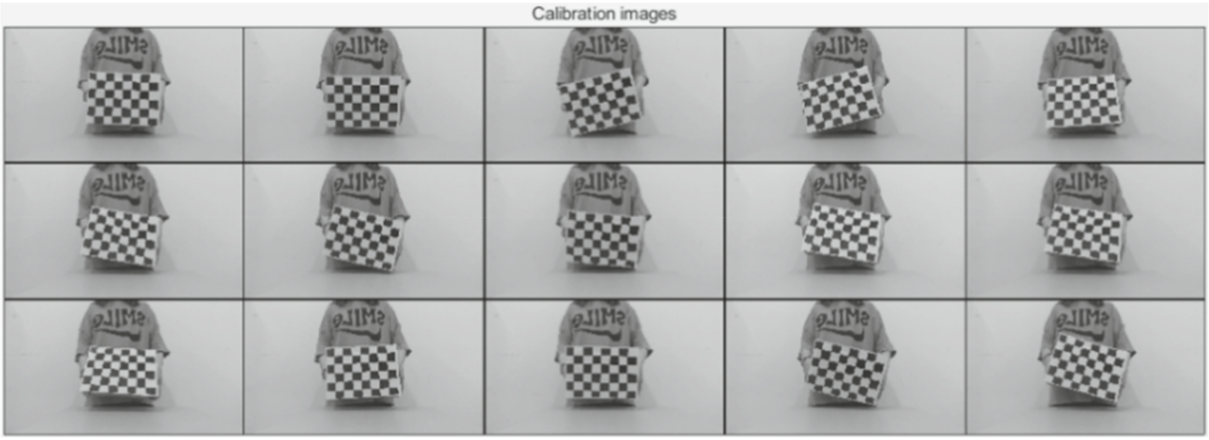
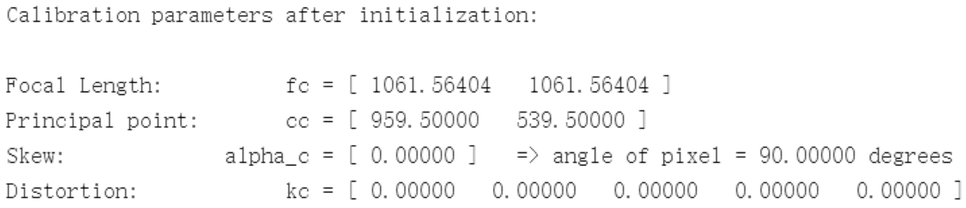


Fig. 2 Chessboard calibration diagram

Fig. 3 Camera internal references before calibration



```
Calibration parameters after initialization:

Focal Length:          fc = [ 1061.56404   1061.56404 ]
Principal point:       cc = [ 959.50000   539.50000 ]
Skew:             alpha_c = [ 0.00000 ]   => angle of pixel = 90.00000 degrees
Distortion:            kc = [ 0.00000   0.00000   0.00000   0.00000   0.00000 ]
```

Table 1 Camera internal references before calibration

Parameter	Numerical value
Focal length	$\left[1061.56404 \;\; 1061.56404\right]$
Principal point	$\left[959.500 \;\; 539.500\right]$
Skew	[0.000]
Distortion	[0.000]

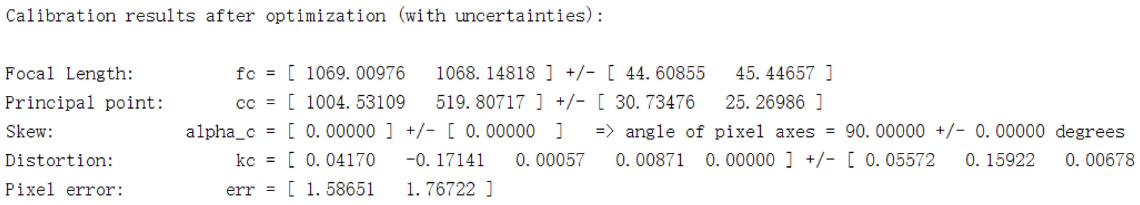
$$Z_c = \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = H\begin{bmatrix} R & t \end{bmatrix}\begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix}. \tag{6}$$

Among them:

$$H = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}. \tag{7}$$

In formula 5, Z_c is the depth value of the pixels, f_x and f_y are the focal length of the camera on the X and Y axes, and c_x and c_y are the coordinates of the optical center, respectively. The camera internal reference is H, which represents the internal characteristics of the camera. $[R\ t]$ is the rotation and displacement of the camera relative to the world coordinate system, i.e., the external parameters.

The chessboard method has good robustness and practicability. Fifteen chessboard images of different positions, angles and postures are taken with black and white rectangular chessboard diagrams as calibration boards. They are calibrated with MATLAB as shown in Fig. 2.



```
Calibration results after optimization (with uncertainties):

Focal Length:          fc = [ 1069.00976   1068.14818 ] +/- [ 44.60855   45.44657 ]
Principal point:       cc = [ 1004.53109   519.80717 ] +/- [ 30.73476   25.26986 ]
Skew:             alpha_c = [ 0.00000 ] +/- [ 0.00000  ]   => angle of pixel axes = 90.00000 +/- 0.00000 degrees
Distortion:            kc = [ 0.04170   -0.17141   0.00057   0.00871   0.00000 ] +/- [ 0.05572   0.15922   0.00678
Pixel error:          err = [ 1.58651   1.76722 ]
```

Fig. 4 Calibrated camera internal reference

Table 2 Calibrated camera internal references

Parameter	Numerical value
Focal length	$\left[1069.00976 \;\; 1068.14818 \right] \pm \left[44.60855 \;\; 45.44657 \right]$
Principal point	$\left[1004.53109 \;\; 519.80717 \right] \pm \left[30.73476 \;\; 25.26986 \right]$
Skew	$[0.00000] \pm [0.00000]$
Distortion	$\left[0.04170 \;\; -0.17141 \;\; 0.00057 \;\; 0.00871 \right] \pm \left[0.05572 \;\; 0.15922 \;\; 0.00678 \;\; 0.00861 \;\; 0.00000 \right]$
Pixel error	$\left[1.58651 \;\; 1.76772 \right]$

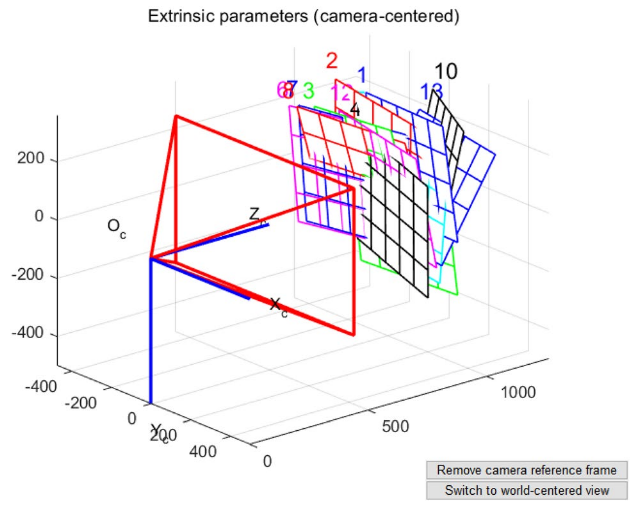


Fig. 5 The position of the calibrated image relative to the camera

The internal reference of the camera is shown in Fig. 3.

The internal parameters of the camera before calibration are arranged as shown in Table 1.

The internal camera of the calibrated camera is shown in Fig. 4.

The calibrated camera parameters are shown in Table 2.

Click Show Extrinsic: display the position of the calibrated image relative to the camera from the camera's perspective (i.e., keep the camera's position and direction unchanged). The position of the calibrated image relative to the camera is shown in Fig. 5.

Among them, 1–15 indicates the position of the calibrated image relative to the camera at different positions, and O_c (X_c, Y_c, Z_c) is the position coordinate of the camera.

Feature extraction and matching

Feature detection is to use computer to extract image information and determine whether each image point belongs to an image feature. The result of feature detection is that the points on the image are divided into different subsets, which often belong to isolated points, continuous curves or continuous regions. The features extracted from different images in the same scene should be the same. Some feature points are extracted from the image and analyzed locally instead of observing the whole image. The requirements of feature points are sufficient, stable and with accurate positioning.

The visual invariance of feature detection is a very important concept, but it is very difficult to solve the problem of scale invariance. To solve this problem, the concept of scale invariant feature is introduced in computer vision. The idea is that not only can objects photographed at any scale detect consistent key points, but each feature point detected corresponds to a scale factor. Ideally, for the same object point with different scales in two images, the ratio between the two scale factors calculated should be equal to the ratio of image scales. In recent years, many scale-invariant features have been proposed. This section introduces one of them, SURF features. Called the SURF "accelerate the steady characteristics" (Speeded Up Robust Feature), it is a kind of method of scale-invariant feature; it provides better robustness detection and description of the child and can be used for target recognition in the field of computer vision or three-dimensional reconstruction, etc. Part of its inspiration comes from the SIFT algorithm, which uses the method of local gradient histogram. The main difference lie in the performance, which reduces the computation time by effectively using the image convolution integral graph. To detect the dimension-invariant feature, the maximum value should be calculated in image and scale space, respectively. As we shall see, they are not only scale-invariant features, but also features with high computational efficiency.

Feature point matching refers to finding the correct matching feature points in two images that need to be registered. The method of feature matching is to first find the feature points with significant features, then describe the two feature points separately, and finally compare the similarity between the two descriptions to determine whether they are the same feature. If scale can be determined before feature description, scale invariance can be achieved. SURF is a scale-invariant feature method and has good robustness of detectors and descriptors. It can be used for object recognition or three-dimensional reconstruction in the field of computer vision. SURF originates from the SIFT algorithm and adopts the method of local gradient histogram. The main difference is that the image convolution integral graph is used to reduce the computational time.

SURF algorithm mainly includes integral image, scale space construction, location and main direction of feature points, and generation of feature descriptors. Integral image refers to calculating the Hessian matrix determinant of each pixel, and calculating the integral image to get the value of each image element.

$$I_{\sum}(x,y)=\sum_{i=0}^{i\le x}\sum_{j=0}^{j\le y}I(x,y). \tag{8}$$

The construction of scale space is to detect the feature points of DOH approximation and get the second-order differential Hessian matrix.

$$H(f(x,y))=\begin{bmatrix}\frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x\partial y}\\ \frac{\partial^2 f}{\partial x\partial y} & \frac{\partial^2 f}{\partial y^2}\end{bmatrix}. \tag{9}$$

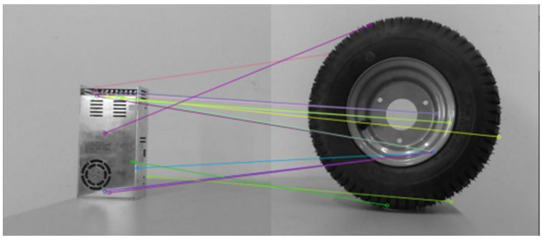
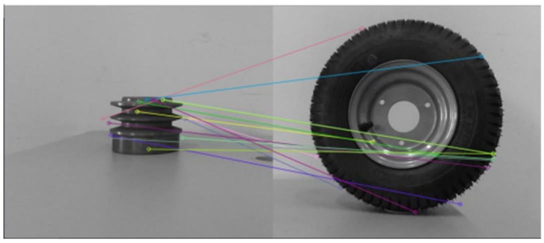
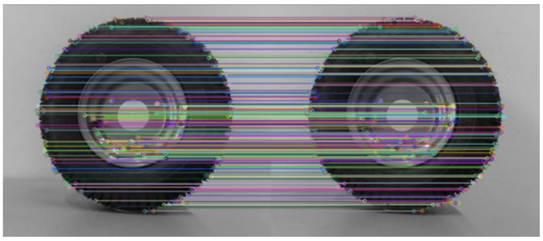
SURF algorithm can process every layer of pyramid image in parallel and build a Gauss pyramid by Hessian matrix response.

OpenCV2.4.10 and Visual Studio 2013 are selected as software support. Firstly, the image information is extracted, and two PNG images generated by RGB-D in each experiment are input. The images are read in black and white mode. Then, the key points of the images are detected by the SURF detector. The minimum Hessian matrix is 400. The eigenvector calculates descriptor and matches descriptor vectors by the violent matching method. Then, the two images are drawn and found. Finally, the matching results are obtained, as shown in Table 3.

By comparing the two images, the similarity between square object and tire matching is only about 20%, while that between cylinder and tire matching is about 30%, and that between cylinder and tire matching is about 80%. When the matching degree is about 80%, the system determines that the part is the target part.

Feature extraction and description are the basis of image processing and computer vision; the image could be affected by noise in practical problems, and the interference of background may also occur. Angle, lighting, scale, translation, rotation, affine changes, such as how to choose reasonable description and image characteristics of the operator, make these

Table 3 Tests matching results

Object Comparison	Key Point Similarity
	21.7%
	30.8%
	79.8%

features not only have good performance of the clock, but also remain unchanged under the above changes, directly determining the effect of image processing based on feature. Based on the invariance theory of computer vision, the study on the invariance of image features has become an important link in image processing, attracting the interest of many researchers.

Table 4 Four tire locations

Point	Coordinate
1	(264, 100, 618)
2	(184, 188, 635)
3	(270, 276, 605)
4	(361, 194, 689)

Three-dimensional positioning of target parts

Positioning is to obtain the position information of the target part in the image. For many parts in the pipeline, the identified target parts are selected by the method of rectangular box. When there are multiple target parts in an image (the number is not fixed), the detection task should locate the target in the image with rectangular frame as much as possible, which is equivalent to the positioning of multiple targets.

According to the imaging principle of the camera, the measuring coordinate system is constructed, and the parameters inside and outside of the color camera and the depth camera are solved.

Implement object recognition is based on open CV. Grayscale processing is carried out for the color pictures collected by Kinect to remove the complex influence of multi-color frequency of color images and make the picture processing more concise. The improved mean filtering algorithm is used to smooth the small noise on the image surface and reduce sharp changes. By means of image binarization, gray value is presented with black and white effect, which makes it no longer involve multi-level value of pixel, simplifies the processing process, and reduces the amount of data processing and compression. Close operation method is used to fill in small voids, smooth boundary and eliminate small and useless voids. According to the object pixel gray value, the method extracts the object contour line, creates a minimum rectangular border based on the set of points and selects the object.

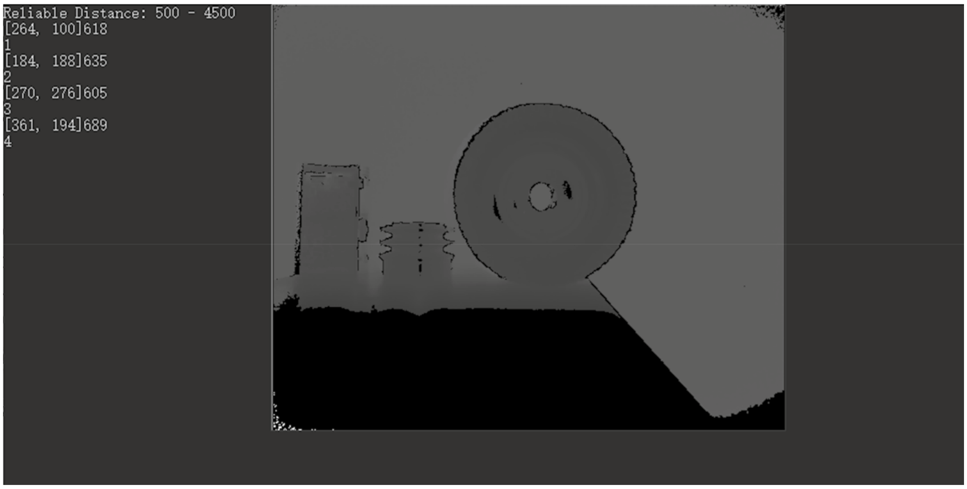
It realizes the function of object recognition and location in a simple environment. According to the collected image, contour analysis, judgment and recognition of the target object, and selection are conducted. We calculate the object center and get the coordinates of the object and output.

After running the corresponding program, the program controls KINECT to obtain the depth image of the target part, as shown in Fig. 6, and determines the depth information of the target part relative to the camera.

Based on the fusion of the RGB image and depth image, the spatial position parameters of the target parts are obtained, and the three-dimensional coordinates of the target parts are also obtained. The tire is the target part to obtain the location information. The four points in the upper left corner of the figure are four different points of the tire, through which the spatial position of the tire is determined. The coordinates of the four points are shown in Table 4.

Summary

One of the cores of multi-part automatic assembly line is the target part identification and positioning system, and the accuracy of identification and positioning affects the efficiency of the automatic assembly line.



```
Reliable Distance: 500 - 4500
[264, 100]618
1
[184, 188]635
2
[270, 276]605
3
[361, 194]689
4
```

Fig. 6 Depth image of the RGB-D camera

Firstly, this paper calibrates the depth camera of Kinect to obtain more accurate internal parameters of the depth camera and improve the accuracy of 3d reconstruction. Then, scale-invariant feature transform (SIFT) algorithm is used to extract image information by computer, and feature description is carried out to determine whether each image point belongs to an image feature and discuss how to extract the feature point in the image. In the end, a simple three-dimensional object reconstruction method is proposed from the aspects of visual perception system construction, object image acquisition, object model construction, object recognition and positioning algorithm in a single background, and the research on object recognition and positioning algorithm in a simple scene is completed.

The experimental results show that the recognition and sorting rate of multi-part automatic assembly line target parts based on the machine vision design recognition and positioning system is significantly higher than the traditional sorting rate.

This research can improve the efficiency of automatic assembly line identification and positioning of multiple parts, and hopefully provide theoretical basis and experimental support for the realization of intelligent automatic assembly line.

Acknowledgements Thanks are due to Dr. Ni for assistance with the experiments and to Dr. Zhu for valuable discussion. This work was supported by the National Natural Science Foundation of China (Grant no. 51405419), Natural Science Foundation of the Jiangsu Higher Education Institutions of China (Grant no. 18KJB460029), Yancheng Institute of Technology Training Program of Innovation and Entrepreneurship for Undergraduates (Grant no. 2019218).

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Kumawat A, Sucheta P (2018) Feature detection and description in remote sensing images using a hybrid feature detector. Procedia Comput Sci 132:277–287
2. Hu J, Peng X, Chengyu Fu (2015) A comparison of feature description algorithms. Optik 126(2):274–278
3. Tang P, Yuxin P (2017) Exploiting distinctive topological constraint of local feature matching for logo image recognition. Neurocomputing 236:113–122
4. Martin F et al (2014) Two different tools for three-dimensional mapping: DE-based scan matching and feature-based loop detection. Robotica 32(01):19–41
5. Shi J, Xuanyin W (2017) A local feature with multiple line descriptors and its speeded-up matching algorithm. Comput Vis Image Underst 162:57–70
6. Yibo L, Li J (2011) Harris corner detection algorithm based on improved contourlet transform. Procedia Eng 15:2239–2243
7. Deshpande PD, Prachi M, Anil ST (2019) Accuracy enhancement of biometric recognition using iterative weights optimization algorithm. Eurasip J Inf Secur 2019:1
8. Endres F et al (2014) 3-D mapping with an RGB-D camera. IEEE Trans Robot 30(1):177–187
9. Guo Y et al (2016) A comprehensive performance evaluation of 3D local feature descriptors. Int J Comput Vis 116(1):66–89
10. Salti S, Federico T, Di Luigi S (2014) SHOT: unique signatures of histograms for surface and texture description. Comput Vis Image Underst 125:251–264
11. Liu W et al (2019) Design of chassis structure and control system for indoor intelligent vehicle. In: Proceedings of the 2019 international conference on artificial intelligence and advanced manufacturing. Association for Computing Machinery, New York
12. Mikolajczyk K, Schmid C (2005) A performance evaluation of local descriptors. IEEE Trans Pattern Anal Mach Intell 27(10):1615–1630
13. Zhang X, Wei L (2019) Feature extraction method of indoor structured environment based on two-dimensional LiDAR. In: 2019 3rd international conference on robotics and automation sciences (ICRAS). IEEE, Wuhan
14. Shi B, Zhang Q, Xu H (2019) A geometrical-information-assisted approach for local feature matching. Math Probl Eng 2019:1409672
15. Wang Z, Guohui T (2020) Integrating manifold ranking with boundary expansion and corners clustering for saliency detection of home scene. Neurocomputing 379:182–196
16. Murartal R, Tardos JD (2017) ORB-SLAM2: an open-source SLAM system for monocular, stereo, and RGB-D cameras. IEEE Trans Robot 33(5):1255–1262
17. Hodan T et al (2017) T-LESS: an RGB-D dataset for 6D pose estimation of texture-less objects. In: 2017 IEEE winter conference on applications of computer vision (WACV). IEEE, Santa Rosa, CA
18. Yew ZJ, Lee GH (2018) 3DFeat-Net: weakly supervised local 3D features for point cloud registration. In: Ferrari V, Hebert M, Sminchisescu C, Weiss Y (eds) Computer Vision – ECCV 2018. Springer, Cham

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