Journal of Computing and Electronic Information Management

ISSN: 2413-1660

A dynamic target recognition method for monocular manipulator

Xinyue Niu[1, a], Jiexin Pu[1, b] and Chi Zhang[1, c]
College of Information Engineering, Henan University of Science & Techology, Luoyang 471003, China.

[a]nxy_sherl @163.com, [b]pujiexin@163.com, [c]zhangchi12306@163.com

Abstract: In this paper, a new SIFT feature point extraction strategy and a low-dimensional feature description scheme are proposed. At the same time, the ViBe method is used to detect moving targets, which improves the integrity of the detected targets. Experiments show that the method proposed in this paper shortens the extraction time of SIFT feature points, and at the same time guarantees the number of pairs of matching points and the matching accuracy.

Keywords: SIFT algorithm, ViBE, Image matching, Monocular manipulator.

1. Introduction

In industrial production, vision is an important source of information for industrial robots to perceive the external environment [1]. Vision-guided positioning technology offers many advantages such as non-contact, high efficiency and fast dynamic response, which greatly improves the flexibility and operational feasibility of industrial robots [2,3]. Therefore, the applications of visually-guided robotic system with only one eye-in-hand camera are extensive, from pick-and-place and peg-in-hole tasks to high precision autonomous assembly systems [4-6].

Target detection is one of the core issues of monocular vision positioning [7], and the image matching is the key technology to achieve it. Therefore, the final effect of visual positioning can be directly affected by the matching algorithm [8]. Matching algorithms based on image feature are usually used to achieve image matching. SURF [9-10], ORB [11,12], SIFT [13-16] are some widely known methods. Among them, SIFT (Scale Invariant Feature Transform) [17-19] is a local feature operator based on image scaling, rotation, affine transformation, which can maintain the invariant of image and suppress a certain degree of noise and the effect of viewing angle changes. The SIFT algorithm is a better solution to image matching in fixed backgrounds or

some scenarios with low real-time requirement. However, this algorithm is of high computational complexity because of the large amount of convolution calculations in the detection process of feature points, which results in a long matching time. Therefore, it is not practical in a monocular eye-in-hand system where the real-time requirement is very high.

In response to the above deficiencies, many scholars have invested in research to improve the SIFT algorithm. Literature [20,21] employs principal component analysis (PCA), which uses low-dimensional subspaces to represent high-dimensional data to compress vector dimensions. Literature [22] proposed the gradient location-orientation histogram (GLOH) algorithm, using concentric circles as feature descriptors, and then reducing the dimension by principal component analysis.（It should be noting that the above methods）All of the above methods effectively utilize the advantages of PCA technology, but increase the workload of training the projection matrix, and the universality of the effects of such methods is limited by the type of training pictures. In [23], the 96-dimensional descriptor is introduced, and the Canny operator is used to eliminate the edge response point, which makes the extreme point more accurate. In [24]and [25], the feature descriptor is constructed by using a circle to obtain a 64-dimensional feature description vector, which enhances the rotation invariance of the feature descriptor. Both of the two algorithms improved in real-time performance, but owing to the high-demanding environment for acquiring images, it is not suitable for the case where the disturbance of working conditions in industrial production is uncertain. In [26], a "nested box"-shaped double square neighborhood window is used to divide the feature point neighborhood, and the 32-dimensional feature vector is established to represent each feature point, which realizes large dimensionality reduction and shortens the matching time. Due to the advantages of Harris corner detection algorithm, such as simple calculation and unaffected by illumination, rotation and noise, the combination of Harris corner detection algorithm and the SIFT feature description has been proposed in[26-28]. Fast and robust image registration is achieved in these methods, however, when the unknown image rotation changes greatly, the matching rate decreases.

Based on the above discussion, a feature point detection method based on SIFT-harris is proposed is proposed In this paper, and the Gaussian circular window is used to create a 32-dimensional feature descriptor, the ViBE algorithm is adopted to extract moving targets. A fast and accurate recognition of moving targets in dynamic scenes is realized.

2. Preliminaries and Problem Formulation

Compared with the target recognition in static scenes, the recognition and positioning

process of the workpiece in the motion background contains more information, which makes the recognition of the moving target more challenging. For existing methods[30,31], the target detection and recognition in a dynamic background usually contains the following steps: First, a global motion parameter model is established and the model parameters are accurately estimated. At the same time, it is crucial to choose a feature point selection and matching method that is easy to identify, easy to extract, and insensitive to noise. Then, the background compensation is performed based on the estimation of global motion, and the moving target detection problem in the motion background is transformed into the moving target detection in the static background. Finally, the Temporal Difference method is used to detect the moving target. The specific process is shown in Figure 1, and further analysis is given for these steps.

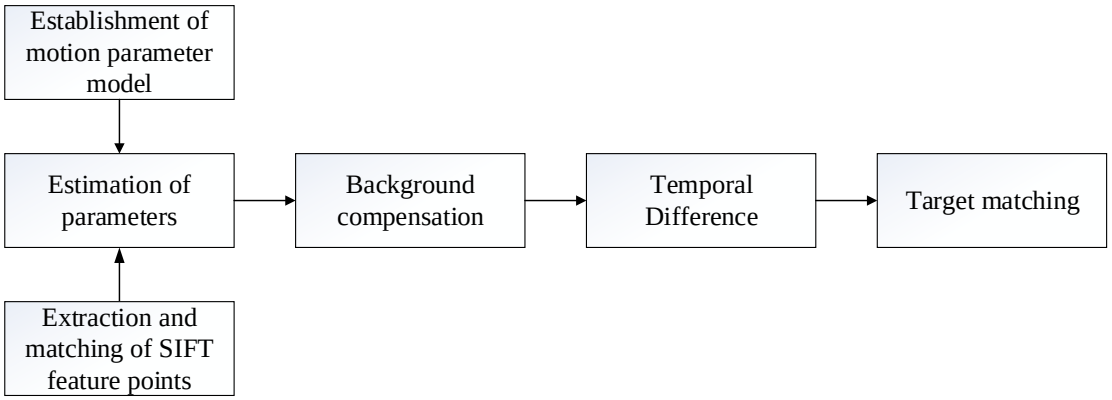

Fig. 1 Flow chart of general method for target detection and recognition in dynamic background

2.1 Global Motion Parameter Model

The global motion parameter model is usually modeled by a 6-parameter or 8-parameter affine transformation model. Although the affine model can describe linear transformations such as translation, rotation and scaling of planar images, this model can only map plane images in parallel, which means that it requires the target scene to be far enough away from the camera so that the target scene is considered as a plane. Therefore, the utilization of this model may be limited, especially for the application of monocular robot workpiece recognition, which forced us to find a more suitable method to establish the motion parameter model.

2.2 Extraction of Image Feature Points

As we all know, because of the good invariance of rotation, scaling, changing viewpoint and affine distortion, the SIFT algorithm is widely used for image matching. Generally, the SIFT feature extraction consists of four steps: (1) Detection of extreme

points in the scale space. (2) Accurate localization of keypoints. (3) Direction determination of keypoints. (4) Generation of feature vectors. The general flow chart of this algorithm is given as follows.

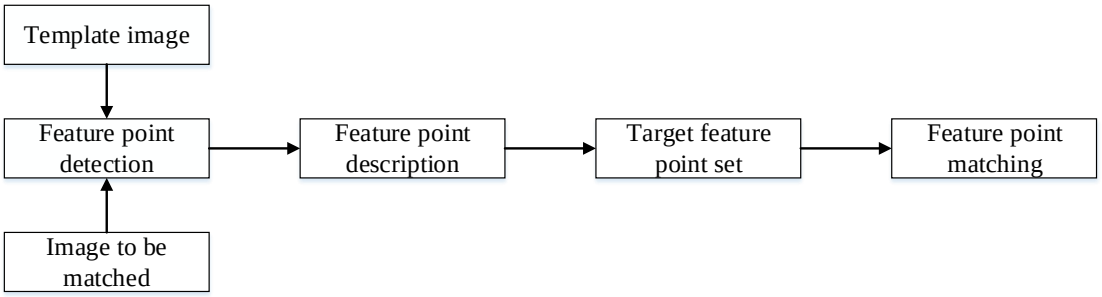

Fig. 2 Flow chart of SIFT Algorithm

Firstly, in the process of searching for extreme points in the scale space, the Gaussian differential function is mainly used to find the points of interest that are not sensitive to the scale change, which is the basis for detecting the invariant features. For a two-dimensional image, different scale spaces can be defined as the convolution of the image with the Gaussian kernel.

Moreover, for the purpose to efficiently detect stable feature points in the scale space, Lowe uses the differential Gaussian DoG extremum as the basis for judgment. Let k be the scale factor between two adjacent scales, the DoG operator is defined as follows:

$$D(\mathrm{x},\mathrm{y},\sigma) = (G(\mathrm{x},\mathrm{y},\mathrm{k}\sigma) - G(\mathrm{x},\mathrm{y},\sigma)) * I(x,y) = L(\mathrm{x},\mathrm{y},\mathrm{k}\sigma) - L(\mathrm{x},\mathrm{y},\sigma) \quad (1)$$

The extreme point is determined by comparing the DoG value of the detection point with that of the pixel within the same scale (adjacent scale). Obviously, multiple calculations of the Gaussian kernel function and the image are required by the algorithm mentioned above, so the large amount of computation and poor real-time performance are the problems we have to face.

2.3 Description of Feature Points

During the description of the key points, Lowe rotates the coordinate axis to the main direction of the feature point to ensure the rotation invariance of the feature point. 4*4 sub-regions are divided in the neighborhood centered on the feature points, and the histogram is used to calculate the gradient value of each sub-region in 8 directions, and 16*8=128-dimensional descriptors are obtained. When matching feature points, it is necessary to calculate the Euclidean distances of all corresponding dimensions of the feature descriptors, and finally the matching degree of the two feature points of different image frames is obtained. Although constructing a 128-dimensional feature vector guarantees the stability of the algorithm performance, it brings a huge amount of computation. Naturally, we consider that whether the computational cost can be reduced by decreasing the feature vector dimension directly while ensuring the

matching performance.

2.4 Detection of Moving Targets

When the Temporal Difference method is used to detect a moving target, "holes" may appear in the detection result. In this case, when the moving target is small, the problem can be solved by simple threshold processing and morphological operations. However, when the target is large, the target detected by the method is usually incomplete, and sometimes even a ghost phenomenon may occur, in this case, the algorithm fails to extract the moving target. Therefore, to improve the accuracy of moving target detection, new methods are needed to subtract background.

Our research is inspired by the discussion of four aspects mentioned above. A dynamic target recognition method based on the improved SIFT method for monocular manipulators is proposed in this paper.

3. Design and Implementation of New Methods

3.1 Establishment of Global Parameter Model

A planar homography transformation is used to establish a parametric model to accurately describe the motion of the camera. The planar homography transformation is used to establish a parametric model to accurately describe the motion of the camera, which can correlate the position of the feature point set in the previous image frame to be matched with that of the current image frame. The homography matrix is defined as:

$$H=\begin{pmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{pmatrix}=\begin{pmatrix} h_1^T \\ h_2^T \\ h_3^T \end{pmatrix} \tag{2}$$

Let $h_{33}=1$, the planar homography matrix has only 8 degrees of freedom. Supposing $\alpha=(\mathrm{x},y,1)^T$ and $\beta=(\mathrm{u},v,1)^T$ are the homogeneous coordinates corresponding to the matching points, then α can be transformed to β by the homography matrix:

$$\beta=H\alpha \tag{3}$$

The matrix H includes transformations such as translation, rotation, and scaling between two adjacent frames. Expand (3), we get:

$$\begin{pmatrix} u \\ v \\ 1 \end{pmatrix}=\begin{pmatrix} h_1^T\alpha \\ h_2^T\alpha \\ h_3^T\alpha \end{pmatrix}=\begin{pmatrix} h_1^T\alpha / h_3^T\alpha \\ h_2^T\alpha / h_3^T\alpha \\ 1 \end{pmatrix} \tag{4}$$

That is:

$$\begin{cases} h_1^T\alpha-u(h_3^T\alpha)=0 \\ h_2^T\alpha-v(h_3^T\alpha)=0 \end{cases} \tag{5}$$

Substituting $h_i^T = (h_{i1}, h_{i2}, h_{i3})^T$ into the above equation, we obtain the eight-element one-time equation w.r.t. h_{ij}:

$$\begin{cases} h_{11}x + h_{12}y + h_{13} - uh_{31} - uh_{32}y = u \\ h_{21}x + h_{22}y + h_{23} - vh_{31} - vh_{32}y = v \end{cases} \quad (6)$$

Theoretically, a plane homography matrix of 8 degrees of freedom can be calculated by four pairs of matching points. In fact, the affine transformation can be understood as a special case when the homography matrix element h_{31}=h_{32}=0 in the planar homography transformation.

A homography transformation is used to map the previous frame image to the current frame, and the gray value of the pixel at the non-integer position is obtained by bilinear interpolation, which is used for global motion compensation. The homography transformation model can be employed to describe the mapping relationship between plane and plane in 3D space. It is more general than the affine model, and can accurately describe the motion of the camera while reducing the computational complexity.

3.2 Extraction and Description of the Feature Point

The extraction of corner spots of the Harris operator[32] is determined by equations (7) and (8).

$$R = \det M - k(traceM)^2 \quad (7)$$

$$M = \begin{bmatrix} I_x^2 & I_x I_y \\ I_x I_y & I_y^2 \end{bmatrix} \quad (8)$$

Where I_x is the derivative of the point (x, y) in the x direction in the image; I_y is the derivative of the point (x, y) in the y direction in the image, a and b are the two eigenvalues of M. In the formula (1), $\det M$ represents the determinant of M, which is equal to the sum of a and b. $traceM$ represents the trace of M, which is equal to the product of a and b. k is a constant (typically 0.04-0.06). When the value of R is greater than a certain threshold value and a local extremum is obtained within a certain neighborhood, it is marked as a corner point.

In this paper, a Gaussian circular window is used to establish a 32-dimensional feature description vector for selected feature points. Figure 3 is a schematic diagram of establishing the neighborhood of feature points. In the algorithm, the feature point is used as the origin, θ is the polar angle to construct a two-dimensional coordinate system, which is divided into 32 sub-regions by the feature point neighborhood, and the feature vectors $TH(x, y) = (HR_1, HR_2 ... HR_{32})$ are standardized, then the 32-dimensional feature descriptor is created for each feature point. Moreover, after being normalized, the feature descriptor has good adaptability to the unknown image that

has been fuzzy transformed.

$$HR_i = \frac{1}{NR_i} \sum_{(x,y)\in R_i} trace(x,y) \tag{9}$$

Where NR_i represents the sum of the pixel points contained in each sub-area, and (x,y) is the two-dimensional coordinates of the pixel points.

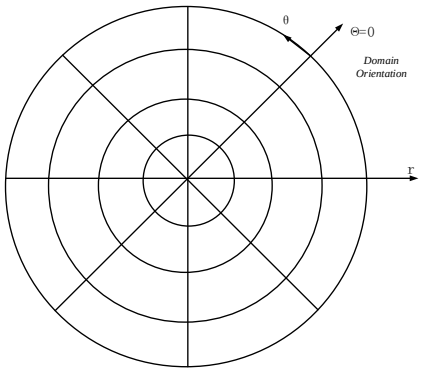

Fig. 3 Schematic Diagram of Establishing a Feature Point Neighborhood

For the purpose to make the feature descriptors be invariance to small angle, a principal orientation, also known as a reference orientation, must be determined for the feature points according to the local image features of the feature point neighborhood. The formula for calculating the polar angle is as follows:

$$\theta = a\tan\left(\frac{y-y_c}{x-x_c}\right) \tag{10}$$

Where x_c and y_c are the coordinates of the feature points.

3.3 Estimation of Motion Parameters

The accuracy of the estimation of the global motion parameters determines the effect of the background compensation, which in turn affects the accuracy of the target detection. After the matching feature point pairs are obtained, the matching SIFT feature point pairs are substituted into equation (9) to estimate the motion parameters of the background model. That is to say, when using the planar homography transformation to establish the global parameter model, 8 motion parameters can be obtained based on only 4 pairs of matched SIFT feature points. Generally, the number of matching feature point pairs is often much larger than 4, so the least square method is used, and the optimal function matching of the data is obtained by minimizing the sum of the squares of the errors. Thereby, the optimal value of the objective function is obtained, and the optimal solution of the matrix vector and the motion parameters of the camera are obtained.

3.3 Moving Target Detection Based on ViBE Method

The moving target detection in the dynamic scene is performed after the matching

SIFT feature point pairs are found. Since the 8-parameter planar homography transformation is used, the homography matrix needs to be solved to obtain the values of h11, h12, h13, h21, h22, h23, h31 and h32. Where h13 and h23 represent translation vectors, and h11, h12, h21, h22 are synthetic matrix representations of rotation, scaling, and shear.

4. Experimental Comparison and Analysis

The experimental environment is Intel (R) Core (TM) i5 -2450M CPU@2.5GHZ2.50GHZ processor with 4.00GB memory, the simulation platform is Matlab2010b, and the operating system is Windows7. Several template images are collected from an experimental platform that simulates an industrial production environment. Subsequently, these template images are rotated, scaled and fuzzy transformed, and the processed images are used for simulation experiments to evaluate the performance of the proposed algorithm.

4.1 Matching Rate

The template images acquired by the monocular vision are rotated, blurred, and scaled to obtain a series of images for matching tests. The SIFT method and the improved SIFT method proposed in this paper are used to obtain the feature matching results. The number of feature points extracted on the image to be matched and the number of feature points that can match the template image are counted in the matching result. Then, according to the definition of the matching rate in [33], the corresponding matching rate is obtained, the performance of the two image matching algorithms is compared. The experimental results are shown in Figures 4-5 below.

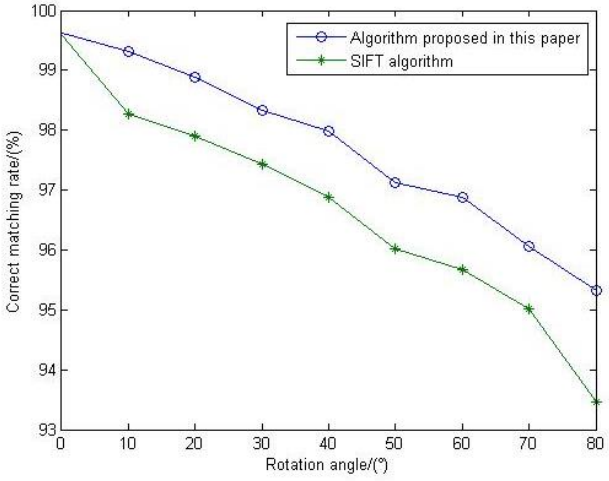

Fig. 4 Comparsion of Rotation Transformation

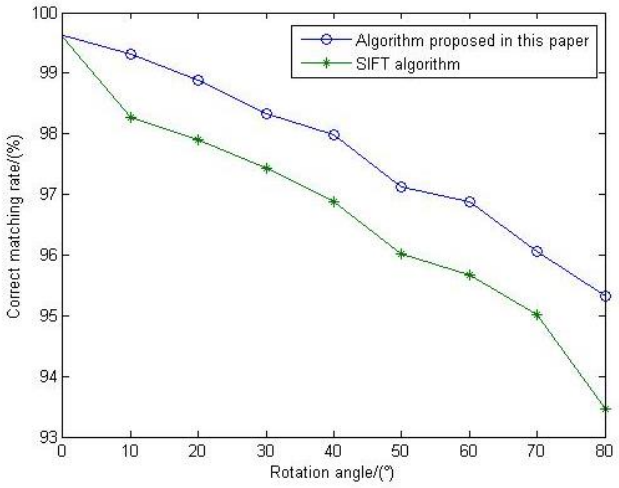

Fig. 5 Comparsion of Pixel Transformation

It can be seen from Figure 4 that the matching rate obtained by the two methods does not change greatly with the angle, and it can remain stable. This means that the proposed method inherits the characteristics of the SIFT operator for angle invariance. Figure 5 shows that the matching rate of this method is higher than that of the sift method under different pixels. However, the method proposed in this paper reduces the calculation amount of the whole process, ensures the matching rate, improves the efficiency of the algorithm, and is more in line with the requirements of workpiece recognition under dynamic background.

4.2 Establishment of Motion Model

(a) Frame 35 (b) Frame 45

Fig. 6 Two Frames Captured by the Camera During the Movement of the Workpiece

After the feature extraction is performed by using the SIFT algorithm and the improved SIFT algorithm proposed in this paper respectively, the eight motion parameters of the test video in Figure 7 are calculated, and the values obtained are shown in Table 1.

Table 1 Transformation Model Parameters Obtained Before and After Algorithm Improvement

	h_{11}	h_{12}	h_{13}	h_{21}	h_{22}	h_{23}	h_{31}	h_{32}
SIFT Algorithm	1.7798	0.8369	4.9723	-0.0096	0	-3.4921	2.37	1.387
Algorithm proposed in this paper	1.7845	0.7993	5.4936	0.0011	0.0032	-2.9967	3.01	1.269

5. Conclusion

This paper presents a dynamic target recognition method for monocular manipulators based on improved sift method. Through the analysis of the traditional SIFT algorithm, it is found that there are some problems such as large amount of computation and slow speed in the extraction of feature points and the generation of feature descriptors. Therefore, in this paper, a new SIFT feature point extraction strategy and a low-dimensional feature description scheme are proposed. At the same time, the ViBe method is used to detect moving targets, which improves the integrity of the detected targets. Experiments show that the method proposed in this paper shortens the extraction time of SIFT feature points, and at the same time guarantees the number of pairs of matching points and the matching accuracy.

References

[1] M. Ito, M. Shibata. Visual tracking of a Hand–Eye robot for a moving target object with multiple feature Points: translational motion compensation approach[J]. Advanced Robotics, 2011, 25(3-4)：355–369.

[2] A. Agrawal, Y. Sun, J. Barnwell. Vision-guided robot system for picking objects by casting shadows[J]. The International Journal of Robotics Research, 2010, 29(2-3):155–173.

[3] T. Kroger, J. Padial. Simple and robust visual servo control of robot arms using an on-line trajectory generator[C]. 2012 IEEE International Conference on Robotics and Automation, 2012:4862-4869.

[4] P. Corke. Robotics, vision and control: fundamental algorithms in MATLAB[J]. Industrial Robot, 2017, 39(6):75-85.

[5] E. Royer, M. Lhuillier, M. Dhome, et al. Monocular vision for mobile robot localization and autonomous navigation[J]. International Journal of Computer Vision, 2007:74(3): 237–260.

[6] H. Wang, Y. Liu. A new approach to dynamic eye-in-hand visual tracking using nonlinear observers[J]. IEEE Transactions on Mechatronics, 2011, 16(2):387-394.

[7] T. Y. Wang, W. B. Dong, Z. Y. Wang. Position and orientation measurement system based on mnocular vision and fixed target[J]. Infrared and Laser Engineering, 2017, 46(4):1-8.

[8] W. Xie, Z. Li, X. Tu, et al. Switching control of image-based visual servoing with laser pointer in robotic manufacturing systems[J]. IEEE Transactions on Industrial Electronics, 2009, 56(2):520-529.

[9] W.-C. Chang. Robotic assembly of smartphone back shells with eye-in-hand visual servoing[J]. Robotics and Computer-Integrated Manufacturing, 2018, 50:102–113.

[10]S. Huang, Y. Yamakawa, T. Senoo, et al. Realizing peg-and-hole alignment with one eye-in-hand high-speed camera[C]. In Proceedings of the IEEE International Conference on Advanced Intelligent Mechatronics, 2013:1127–1232.

[11]H. Bay, T. Tuytelaars, L.V. Gool. SURF: speeded up robust features[C]. European Conference on Computer Vision, 2006:404-417.

[12]T. Y. Wang, W. B. Dong, Z. Y. Wang. Position and orientation measurement system based on mnocular vision and fixed target[J]. Infrared and Laser Engineering, 2017, 46(4):1-8.

[13]E. Rublee, V. Rabaud, K. Konolige, et al. ORB: An efficient alternative to SIFT or SURF[C]. 2011 International Conference on Computer Vision, 2011:2564-2571.

[14]R. Mur-Artal, J. Montiel, J. Tardos. ORB-SLAM: a versatile and accurate monocular SLAM system[J]. IEEE Transactions on Robotics, 2015, 31(5):1147-1163.

[15]D.G. Lowe. Object recognition from local scale invariant features[C]. International Conference on Computer Vision, 1999:1150-1157.

[16]D.G. Lowe. Distinctive image features from scale-invariant key-points[J]. International Journal of Computer Vision, 2004, 60(2):91-110.

[17]E. Ahmed, H. David, D. Larry. Non-parametric model for background subtraction[C]. Proceedings of the 6th European Conference on Computer Vision, 2000:751-767.

[18]H. Zhou, Y. Yuan, C. Shi. Object tracking using sift features and mean shift[J]. Computer Vision and Image Understanding, 2009, 113(3):345-352.

[19]B.He, Z. Chen. Determination of the common view field in hybrid vision system and 3D reconstruction method[J]. ROBOT, 2011, 33(5):614-620.

[20]Y.Shireen, Elhabian, M.Khaled, et al. Moving object detection in spatial domain using background removal techniques-state-of-art[J]. Recent Patents on Computer Science, 2008, 1:32-54.

[21]N. Shao, H. G. Li, L. Liu, et al. Stereo vision robot obstacle detection based on the sift[C]. 2010 Second WRI Global Congress on Intelligent Systems, 2010:274-277.

[22]R. Lenz, R. Tsai. Techniques for calibration of the scale factor and image center for high accuracy 3S machine vision metrology[J]. IEEE Trans Pattern Anal Machine Intell, 1988,6(3):323-343

[23]X. Wu, J. Cen, X. Tai. SPCA: A fast dimensionality reduction method for image feature[J]. Journal of Ningbo University, 2005, 18(3):336-339.

[24]K. Mikolajczyk, C. Schmid. A performance evaluation of local descriptors[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(10): 1615-1630.

[25]C. Li, C. Tao, G. Liu, 3D visual SLAM based on multiple interative closest point[J]. Mathematical Problems in Engineering, 2015, 2015:1-11.

[26]C. Zhang, Z. Gong. Improved SIFT feature applied in image matching[J]. Computer Engineering and Applications, 2008, 44(2):95-97.

[27]H. Liu, H. Shen. Image match method based on improved SIFT algorithm[J]. Optoelectronic Technology. 2013, 33(4):249-254.

[28]P. Azad, T. Asfour, R. Dillmann. Combining harris interest points and the SIFT descriptor

for fast scale-invariant object recognition[C]. Intelligent Robots and Systems, 2009:4275-4280.

[29]L. Li, C. Li, X. Zeng, et al. An automatic image registration method based on SIFT and Harris-affine features matching[J]. Journal of Huazhong University of Science and Technology, 2008, 36(8):13-16.

[30]J. Xu, Y. Zhang, H. Zhang. Fast image registration algorithm based on improved Harris-SIFT descriptor[J]. Journal of Electronic Measurement and Instrumentation, 2015, 29(1):48-54.

[31]J. Dou, Q. Qin, Z. Tu. Robust image matching based on the information of SIFT[J]. Optik, 2018, 171: 850-861.

[32]M. Van Droogenbroeck, O. Paquot. Background subtraction: experiments and improvements for ViBe[C]. Computer Vision and Pattern Recognition Workshops, 2012:32-37.

[33]P. Xu, D. Z. Yao, A study on medical image registration by mutual information with pyramid data structure[J]. Computers in Biology and Medicine, 2007, 37:320-327.