Feature Extraction and Matching Algorithm Based on Improved SIFT

Xi Chen[1], Sisi Sun[1], Junying Wu[1], Zihan Zhang[2](✉), Jiao Peng[1], Yanyan Lu[1], and Rukun Liu[1]

[1] State Grid Hebei Information and Telecommunication Branch, Shijiazhuang 050000, China
[2] North China Electric Power University, Baoding 071003, Hebei, China
zhangzihan0321@163.com

Abstract. To solve the problem that the SIFT algorithm used in augmented reality algorithm can not accurately extract matching and the matching efficiency is low, a matching method combining SIFT algorithm, fast explicit diffusion FED and IMU is proposed. Firstly, the input image of the initial matching is processed by a nonlinear filtering operation using FED. Pre-integration is performed by the IMU inertial measurement unit to calculate the pose change of the camera between two frames of images. Then, the SIFT algorithm is adopted for image matching. Finally, the corresponding relationship between the virtual object coordinates registered on the screen and the real space coordinates is determined. With the help of IMU and the nonlinear image fuzzy processing, the proposed method has strong adaptability to images and can obtain the corresponding coordinates of real space in complex environments.

Keywords: SIFT · IMU · Nonlinear filtering · Extraction · Match

1 Introduction

Image feature extraction and matching location are the basic technologies of augmented reality vision. Image matching based on local invariant features is applied to obtain spatial coordinate information, which is to realize the registration of three-dimensional objects into the real environment in augmented reality. Spatial information acquisition mainly includes three steps: extraction of descriptors, description, and matching [2]. At present, the main image feature extraction matching algorithms are the ORB algorithm, SURF algorithm, and SIFT algorithm. The feature points extracted by the ORB algorithm have scale and rotation invariance. The detection efficiency is high, which meets the real-time effect of augmented reality on mobile devices. However, the problem of scale invariance is not considered [3]. Based on the SURF algorithm, the Hessian matrix is used to screen feature points, which improves the detection speed. However, there are still problems, such as poor real-time performance of the system, jitter, and drift of 3D registered virtual objects [4].

Moravec et al. used the corner operator to achieve stereo-vision matching in the early days. Harris et al. improved the Moravec operator on this basis. However, it is sensitive

C. F. Peng et al. (Eds.): EIMT 2023, AHSSEH 8, pp. 162–171, 2023.
https://doi.org/10.2991/978-94-6463-192-0_23

to scale, viewing angle, and illumination changes, and has poor anti-noise ability [5]. Lowe et al. proposed a more stable SIFT (scale in variant feature transform) feature operator, which not only has the invariance of scale, rotation, affine, visual angle and illumination, but also maintains a good matching effect for the motion, occlusion, noise and other factors of the target [6]. Therefore, it has been widely used in image recognition, target tracking, and positioning [7–9]. However, due to the strong randomness of the corresponding feature points of SIFT matching, SIFT can not directly obtain the exact location of the target or provide a more accurate pose in a short time. Thus, a variety of SIFT variants have been proposed.

Document [10] proposes an extraction and matching method based on the simplified descriptor of cross-shaped partition in the field of feature points. Document [11] proposes an extraction and matching algorithm based on the multi-resolution pattern recognition. Document [12] proposes a K nearest neighbor algorithm (KNN) method based on dune tracking, but the number of feature points is small. The matching accuracy is low, and a high error matching rate and the like occur sometimes. In this paper, an image extraction and matching method based on SIFT algorithm is proposed. Through the SIFT algorithm, the Fast Explicit Diffusion (FED) nonlinear filtering, and the Inertial Measurement Unit (IMU), the characteristic position coordinates of the target image are obtained.

2 A SIFT-Based Method for Feature Match and Location

2.1 SIFT Algorithm

SIFT mainly consists 7 steps [6].

1) The input image is smoothed with Gaussian filtering.
2) Construction of the pyramid structure is like this: the image is generally divided into several parts to improve the processing speed. Each part has the same size. Then, a pyramid structure is built on each part so as to save the amount of computation wasted, which is due to the large pyramid structure. Each pyramid is generally divided into 3 to 5 layers, and each layer has a different scale. For example, the bottom layer of the pyramid is the fragment of the original image. The second layer is the Laplace-Gauss convolution transform of the first layer, which blurs the details but reduces the size of the image. Similarly, the third layer is obtained by the Laplace-Gauss transform of the second layer, and the size continues to decrease. When the top layer is reached, a pyramid-shaped model is established. The clarity of each layer is becoming more and more blurred intuitively. Meanwhile, to store the data, the number of faces of the image stored in the layer at the top of the pyramid is higher. There is a down-sampling relationship between the faces between the layers. Therefore, the first face of the 0th layer can be obtained by down-sampling the third face of the first layer. Then, a Gaussian convolution operation similar to that of the 0th layer is performed.
3) Build the DoG pyramid. The Gaussian difference pyramid is constructed through the above pyramid model. Since the construction principle of the Gaussian difference pyramid is to subtract between two faces of the same layer in the original Gaussian pyramid, the number of layers of the Gaussian difference pyramids is the same as that of the original Gaussian pyramid. The number of each layer's faces is 1 less than that of the original Gaussian pyramids.

4) Detect the feature points by using the difference of Gaussian pyramid, wherein in the detection process, if the curvature, gradient and other parameters of the points do not meet the preset threshold number, the points are removed. The feature points in this step are carried out in several faces of each layer. The judgment matrix is a three-dimensional matrix, and the comparison value is mainly the gray value after binarization.
5) Calculate the scale of each feature point.
6) The gradient value and the direction of each feature point are calculated. The points in a matrix region around the feature point are used to describe the feature point. The histogram is used to count the gradient values and find the main direction, which can be more than one.
7) After that accurate gradient value and direction of each feature point are obtained, a feature descriptor through the data was generated, wherein the descriptor has 64-dimensional or 128-dimensional features. It can perform relatively accurate matching on the feature points.

2.2 Nonlinear Multi-scale Spatial Image Processing

To solve the problem of losing the edge and detail information of the matching image, the SIFT feature extraction and matching algorithm is improved in the early stage. The improved algorithm first uses nonlinear filtering instead of linear Gaussian filtering to construct the pyramid scale space, which can better retain the edge and detail information in the process of image blurring. Then, it uses the SIFT feature extraction algorithm to extract feature points in the nonlinear scale space, and applies the rBRIEF algorithm to describe the feature points to get the feature descriptor. Finally, the Hamming distance between the two sets of feature descriptors is calculated [13]. Then, the PROSAC algorithm is adopted to eliminate the mismatched feature point pairs. The construction process of the nonlinear scale space is similar to that of the Gaussian pyramid. A multi-scale image with the same resolution is obtained by performing a nonlinear filtering operation on the image through an FFD (Fast Explicit Diffusion) algorithm [14]. The expression of the diffusion process for nonlinear filtering is as follows:

$$\frac{\partial M}{\partial t} = div(c(x, y, t) \cdot \nabla M) \tag{1}$$

where, M is the brightness of the image, div represents the divergence calculation, ∇ represents the gradient calculation, and $c(x, y, t)$ is the conduction function. The process of solving nonlinear diffusion filtering by FED is as follows:

$$L^{i+1} = L^{i}(I + \tau A(L^{i})) \tag{2}$$

In the formula, I is the unit matrix, $A(L^{i})$ is the matrix of the image in the dimension i, τ is the time step, and L is the definition of the scale space. If a point has the maximum or minimum value among the surrounding eight points and eighteen neighborhood points in the upper and lower scale spaces, it is determined that this point is a feature point of the image in this scale.

The gradient value and the direction of each feature point are calculated. By using the gradient direction distribution characteristics of the pixels in the neighborhood of the feature points, the direction parameters are assigned to each feature point, so that the operator has rotation in-variance.

$$m(x, y) = \sqrt{(L(x+1, y) - L(x-1), y))^2 + (L(x, y+1) - L(x, y-1))^2} \tag{3}$$

$$\theta(x, y) = \arctan\{[L(x, y+1) - L(x, y-1)]/[L(x+1, y) - L(x-1, y)]\} \tag{4}$$

In the formula, $m(x, y)$ is the modulus value of the gradient at (x, y), $\theta(x, y)$ is the direction of the gradient at (x, y). The scale used by L is the scale of each feature point. The feature point is described by the points in a matrix area around the feature point. The SIFT algorithm takes the feature point as the center, selects a 3 * 3 range nearby, constructs a circular area, and then calculates the gradient histogram of this area, thereby constructing a main direction. Finally, the gradient value and direction of each feature point are obtained.

2.3 Inertial Measurement Unit Pre-integration

IMU pre-integration module is widely used in the field of robot navigation, which can provide more accurate pose estimation in a short time, so that the robot can complete more accurate operations [15]. A preintegration model of that IMU is firstly established. A reference coordinate of a camera is set as C. A world coordinate system is set as W. The reference coordinate of the IMU is set as B. An angular velocity w_B and an acceleration a_B of the camera are obtained according to an AR device end integrating the IMU after the camera moves. The rotation rate $\tilde{w}_B$ and the acceleration $\tilde{a}_B$ of a sensor relative to an inertial system are allowed to be measured. The angular velocity rotation rate and acceleration of the sensor are mainly affected by the additive white Gaussian noise η and the slowly varying angular velocity bias b_g of the sensor. First, the angular velocity $\tilde{w}_B$ relative to the inertial sensor is calculated:

$$\tilde{w}_B^i = w_B^i + b_g^i + \eta_g^i \tag{5}$$

In the formula, w_B^i is the angular velocity of the acquisition device when acquiring the i-th frame of acquisition image. b_g^i is the angular velocity deviation parameter, and η_g^i is the white Gaussian noise of the angular velocity of the acquisition device when acquiring the i-th frame of acquisition image. The rotation relationship of the IMU relative to the world coordinate system at the j-th frame image can be obtained according to angular velocity and sensor angular velocity deviation b as follows:

$$R_{WB}^j = R_{WB}^i Exp\left(\left(\tilde{w}_B^i - b_g^i\right)\Delta t\right) \tag{6}$$

where, R_{WB}^i is the rotation matrix of the i-th frame acquisition image, *Exp* is the exponential mapping calculation formula from Lie algebra to Lie group [16], Δt is the time interval between two adjacent frames of images. $\tilde{w}_B^i$ is the angular velocity of the acquisition device relative to the inertial frame when acquiring the i-th frame acquisition image, and b_g^i is the angular velocity deviation parameter of the acquisition device when acquiring the i frame acquisition image.

The rotation relation of the IMU relative to the world coordinate system when the camera collects the ith frame image is expressed as R_{WB}^{i}. The speed relation of the IMU under the world coordinate system when the camera collects the ith frame image is expressed as V_{WB}^{i}. The translation relation of the IMU under the world coordinate systems when the *jth* frame image is collected by the camera can be expressed as T_{WB}^{j}:

$$T_{WB}^{j} = T_{WB}^{i} + V_{WB}^{i}\Delta t + \frac{1}{2}g_W\Delta t^2 + R_{WB}^{i}\left(\tilde{a}_B^{i} - b_a^{i}\right)\Delta t^2 \tag{7}$$

where, $\tilde{a}_B^{i}$ is the acceleration of the acquisition equipment relative to the inertial system when acquiring the *i-th* frame of acquisition image. b_a^{i} is the acceleration deviation parameter of the acquisition equipment when acquiring the *i-th* frame of acquisition image. g_W is the gravity parameter. V_{WB}^{i} is the speed of the acquisition device in the world coordinate system when acquiring the *ith* frame of image. T_{WB}^{i} is the translation matrix of the *ith* frame of image. This formula describes the relationship between the two sets of IMU data.

The acceleration $\tilde{a}_B$ of the sensor relative to the inertial frame is obtained by the rotation matrix R:

$$\tilde{a}_B(t) = R_{WB}^{iT}(t)(a_W(t) - g_w) + b_a(t) + \eta_a(t) \tag{8}$$

where, η_a^{i} is the white Gaussian noise of the acceleration of the acquisition equipment when acquiring the *i-th* frame acquisition image. a_B^{i} is the acceleration of the acquisition equipment when acquiring the *i-th* frame acquisition image. b_a is the acceleration deviation parameter, and g_W is the gravity parameter.

IMU pre-integration and improved SIFT feature point matching are adopted to jointly solve the camera pose. Then, the camera pose change information is used to calculate the transformation relationship between the camera and the real scene. The transformation matrix $S_{4\times 4}$ is obtained. $S_{4\times 4}$ is composed of a rotation matrix R_{WB}^{j} and a translation matrix T_{WB}^{j}, and $S_{4\times 4}$ can convert a world coordinate system into a screen coordinate system and finally register that virtual object in the real world. Assuming that the coordinate of any feature point P in the real space is (X_W, Y_W, Z_W), the homogeneous coordinate is $(X_W, Y_W, Z_W, 1)$. The homogeneous coordinate of the feature point P in the camera coordinate system is $(X_C, Y_C, Z_C, 1)$. Then, the relationship between the two coordinates satisfies:

$$\begin{bmatrix} X_c \\ Y_c \\ Z_c \\ 1 \end{bmatrix} = \begin{bmatrix} R_{WB}^{j} & T_{WB}^{j} \\ 0_{1\times 3} & 1 \end{bmatrix} \begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix} \tag{9}$$

Under the real world coordinate system, the coordinate of the feature point P projected to the screen is (u, v). The translation amounts are p_x and p_y respectively. The focal lengths in the corresponding directions are f_x and f_y. Then, the projection of the feature point P under the camera coordinate system satisfies:

$$u = \frac{f_x X_c}{Z_c} + p_x, v = \frac{f_y Y_c}{Z_c} + P_y \tag{10}$$

Calculate the camera pose of the *jth* frame image according to the rotation matrix R^j_{WB} and the translation matrix T^j_{WB}, and determine the relationship between the virtual coordinates and the real coordinates of the feature points:

$$Z_C\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} f_x & 0 & p_x & 0 \\ 0 & f_y & p_y & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}\begin{bmatrix} X_C \\ Y_C \\ Z_C \\ 1 \end{bmatrix} = \begin{bmatrix} f_x & 0 & p_x & 0 \\ 0 & f_y & p_y & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}\begin{bmatrix} R^j_{WB} & T^j_{WB} \\ 0_{1\times 3} & 1 \end{bmatrix}\begin{bmatrix} X_W \\ Y_W \\ Z_W \\ 1 \end{bmatrix} \tag{11}$$

where, $(X_W, Y_W, Z_W, 1)$ is the real coordinate of the feature point, $(X_C, Y_C, Z_C, 1)$ is the virtual coordinate of the feature point, (u, v) is the coordinate in the captured image of the characteristic point, and p_x is the translation amount in the horizontal direction of the characteristic point in the captured image. p_y is the translation amount in the vertical direction of the characteristics point in the capture image. f_x is the focal length of the feature point in the horizontal direction in the captured image. f_y is the focal length of the feature point in the vertical direction in the captured image. R^j_{WB} is the rotation matrix of the *i-th* frame captured image, and T^j_{WB} is the translation matrix of the *j-th* frame captured image.

After the gradient value and direction of each feature point are calculated by nonlinear filtering, the data is used to generate a feature descriptor (descriptor), which has 64-dimensional or 128-dimensional features. The remaining first feature points are searched in the local area centered on the central first feature point X_{j}. Then, the correct first feature points are matched as the feature points of the acquired image. That is, based on the central first feature point and the plurality of first feature points, the feature points of the captured image are finally obtained. The advantage of this is that the local area search centered on the central first feature point X_{j} narrows the search range.

The coordinates of the central first feature point $X_j(x_j, y_j, z_j)$ are calculated by the following formula:

$$\begin{bmatrix} x_j \\ y_j \\ z_j \end{bmatrix} = R^j_{WB}\begin{bmatrix} x_i \\ y_i \\ z_i \end{bmatrix} + T^j_{WB} \tag{12}$$

where, $X_i(x_i, y_i, z_i)$ is the coordinate of the first feature point X_{j} of the center in the previous frame. R^j_{WB} is the rotation matrix, and T^j_{WB} is the translation matrix.

Based on the rotation matrix, the translation matrix and the virtual coordinates of the feature points, the real coordinates of the feature points are obtained, that is, the real coordinates of the power equipment to be detected are obtained.

The relationship between the virtual coordinates and the real coordinates of the feature points is as follows:

$$Z_C\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} f_x & 0 & p_x & 0 \\ 0 & f_y & p_y & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}\begin{bmatrix} X_C \\ Y_C \\ Z_C \\ 1 \end{bmatrix} = \begin{bmatrix} f_x & 0 & p_x & 0 \\ 0 & f_y & p_y & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}\begin{bmatrix} R^j_{WB} & T^j_{WB} \\ 0_{1\times 3} & 1 \end{bmatrix}\begin{bmatrix} X_W \\ Y_W \\ Z_W \\ 1 \end{bmatrix} \tag{13}$$

In the formula, $(X_W, Y_W, Z_W, 1)$ is the real coordinate of the feature point, $(X_C, Y_C, Z_C, 1)$ is the virtual coordinate of the feature point, (u, v) is the coordinate in the captured image of the characteristic point, p_x is the translation amount in the horizontal direction of the characteristic point in the captured image. p_y is the translation amount in the vertical direction of the characteristics point in the capture image, and f_x is the focal length in the horizontal direction. f_y is the focal length of the feature point in the vertical direction in the captured image. R^j_{WB} is the rotation matrix of the *i-th* frame captured image, and T^j_{WB} is the translation matrix of the *j-th* frame captured image.

Under the real world coordinate system, if the coordinate of the feature point in the collected image is (u, v), the translation amounts are p_x and p_y. The focal lengths in the corresponding directions are f_x and f_y, and the homogeneous coordinate of the feature point is $(X_C, Y_C, Z_C, 1)$, then:

$$u = \frac{f_x X_c}{Z_c} + p_x, v = \frac{f_y Y_c}{Z_c} + P_y \tag{14}$$

During the matching of the SIFT feature points, the pose change of a camera between two frames of images can be calculated through an IMU pre-integration module. The possible position X_2 of a feature point X_1 in the previous frame of image in the next frame of image is obtained, so that the global search and matching of the SIFT feature points are converted into local area search and matching by taking X_2 as a center. The matching iteration times are reduced, thereby improving the matching precision of the feature points and better eliminating mismatching points.

2.4 Design of the Algorithm in This Paper

Because in the actual production application, the perception of spatial coordinates must be captured from the first frame of image translation position. Then, the angle and coordinates of these special feature points in the case of spatial translation are required. In this paper, the SIFT algorithm, fast explicit diffusion FED nonlinear filtering, and IMU will be combined to calculate the coordinates of the target image in the complex environment according to two adjacent frames of images in space. The process is indicated in Fig. 1 as follows:

The first step is to obtain the velocity, angular velocity, and acceleration in the world coordinate system according to the obtained space target image and the IMU of the space image acquisition device.

The second step is to obtain a rotation matrix of the collected image relative to the world coordinate system and a translation matrix in the world coordinate system based on the velocity, the angular velocity, and the acceleration under the world coordinate obtained by the IMU of the device when the space image is collected.

The third step is to perform nonlinear smoothing processing on the acquired image based on the FED and to extract and match by using the SIFT feature based on the acquired image after the smoothing processing obtaining a virtual coordinate of the acquired image.

The fourth step is to combine the IMU pre-integration result and the SIFT characteristic points extracted from the smoothed image to obtain the real coordinate of the

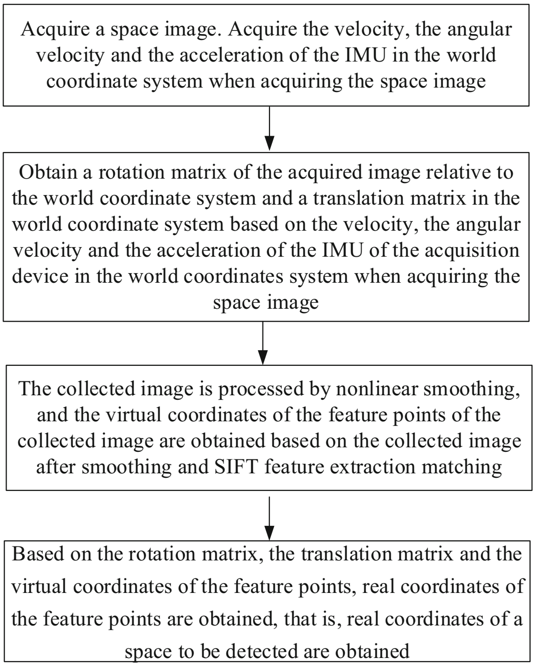
Acquire a space image. Acquire the velocity, the angular velocity and the acceleration of the IMU in the world coordinate system when acquiring the space image
Obtain a rotation matrix of the acquired image relative to the world coordinate system and a translation matrix in the world coordinate system based on the velocity, the angular velocity and the acceleration of the IMU of the acquisition device in the world coordinates system when acquiring the space image
The collected image is processed by nonlinear smoothing, and the virtual coordinates of the feature points of the collected image are obtained based on the collected image after smoothing and SIFT feature extraction matching
Based on the rotation matrix, the translation matrix and the virtual coordinates of the feature points, real coordinates of the feature points are obtained, that is, real coordinates of a space to be detected are obtained

Fig. 1. The algorithm process

space to be detected, which is based on the camera pose transformation obtained by the rotation matrix and the translation matrix in the second step.

3 Conclusion

In this paper, an improved extraction and the matching space positioning algorithm is proposed, which adopts the SIFT algorithm based on image feature point extraction to match the image. To meet the requirements of the camera to accurately locate the target and capture the image, the FED fast explicit diffusion is used to replace the Gaussian pyramid in the early stage of the SIFT algorithm, which is to improve the image smoothing and blurring processing. Much more accurate feature point extraction is obtained to better retain the edge and detail information. IMU is applied for pre-integration to assist the SIFT feature point extraction and matching, which reduces the search range of SIFT algorithm feature points. It also reduces the number of matching iterations and improves the efficiency of feature extraction and matching. Finally, the parameters of the space coordinates were calculated by combining the coordinate model. The correspondence between the virtual camera coordinates and the real space coordinates was also determined. This method has a good spatial localization effect in the case of occlusion, rotation, and other changes in the target image.

Acknowledgment. This paper was supported by the Science and Technology Project of State Grid Hebei Information and Telecommunication Branch: Research projects on key technologies and typical application scenarios based on "5G+AR+AI" (No. SGHEXT00SJJS2100036).

References

1. Luo L , Jun-Qin L I . Realization of Augmented Reality System Based on Natural Features[J]. Computer Knowledge and Technology, 2015.
2. Guo Siyu, Kong Yaguang, Zhang Xufang. Corner detection algorithm based on Hough transform. Journal of Instrumentation, 2008, 29 (11): 6.
3. Rublee E, Rabaud V, Konolige K, et al. ORB: An efficient alternative to SIFT or SURF[C]. In: Proc.of the 13th International Conference on Computer Vision. IEEE Computer Society, Barcelona, Spain,2011:2564–2571.
4. Zhou H, Yuan Y, Shi C. Object tracking using SIFT features and mean shift [J]. Computer Vision &Image Understanding, 2009, 113(3): 345-352.
5. Moravec H P . Obstacle Avoidance and Navigation in the Real World by a Seeing Robot Rover. Stanford University. 1980.
6. Lowe D G . Lowe, D.G.: Distinctive Image Features from Scale-Invariant Key-points. Int. J. Comput. Vision 60(2), 91–110[J]. International Journal of Computer Vision, 2004, 60(2).
7. Wang M, Zhang J , Deng K , et al. Combining Optimized SAR-SIFT Features and RD Model for Multisource SAR Image Registration[J]. IEEE Transactions on Geoscience and Remote Sensing, 2022(60-).
8. Alonso-Fernandez F, Tome-Gonzalez P, Ruiz-Albacete V, et al. Iris Recognition Based on SIFT Features[J]. 2021.
9. Parente L, Chandler J H, Dixon N. Automated Registration of SfM-MVS Multitemporal Datasets Using Terrestrial and Oblique Aerial Images[J]. The Photogrammetric Record, 2021.
10. Tan Guangxing, Zhang Lun. Image feature matching algorithm based on improved SIFT. Journal of Guangxi Institute of Technology, 2022 (002): 033.
11. Xiao Yingnan and Sun Shuyu. Design of high-precision obstacle avoidance algorithm for UAV based on improved SIFT image matching [J]. Mechanical Manufacturing and Automation, 2022, 51 (1): 237-240.
12. Tang Yingfu, Wang Zhongjing, Zhang Zixiong. Dune image registration based on improved SIFT and SURF algorithms. Journal of Tsinghua University (Natural Science), 2021 (002): 061.
13. Ren Jie, Zhou Yu, Yu Yao, et al. Implementation of AR real-time system based on ORB natural features. Computer Application Research, 2012, 29 (9): 3594-3596.
14. Grewenig S, Weickert J, Bruhn A. From box filtering to fast explicit diffusion[C]//Joint Pattern Recognition Symposium. Springer, Berlin, Heidelberg, 2010: 533-542.
15. Ban Chao, Ren Guoying, Wang Binrui, et al. Research on adaptive EKF measurement algorithm for robot attitude based on IMU. Journal of Instrumentation, 2020 (2): 7.
16. Qin T, Shen S . Robust initialization of monocular visual-inertial estimation on aerial robots[C]// IEEE/RSJ International Conference on Intelligent Robots & Systems. IEEE, 2017:4225–4232.