An Improved Indoor Image Registration Algorithm Based on Shallow Convolutional Neural Network Descriptor

Yun Gong and Mengjia Yang(✉)

Xi'an University of Science and Technology, Xi'an 710054, China
2504853112@qq.com

Abstract. At present, the application demand of indoor simultaneous localization and mapping (SLAM) technology increases greatly, among which, image matching is the most basic and critical content. Compared with traditional image registration, indoor image registration has higher requirements on the real-time and robustness of the algorithm. The shallow convolutional neural network is a deep machine learning model based on supervised learning with the characteristics of centralized and automatic learning from data. Aiming at the problems of slow processing and strong rotation failure of feature descriptors in traditional registration algorithms, this paper proposed an improved algorithm of local feature descriptor of triple-sample shallow convolutional neural network, which has strong feature expression ability. In addition, the performance of our improved algorithm was compared with that of three traditional algorithms (SIFT, ORB and SURF) in rotation change of indoor image matching. The results show that the improved algorithm performs better than the other three traditional methods and has a certain antagonistic effect on image rotation.

Keywords: Feature detection · Feature matching · Neural network · Descriptor

1 Introduction

Feature matching refers to a process of seeking common connection points between two images with overlapping areas, which provides a basic support for subsequent data applications. Feature matching is a crucial step in computer vision visualization. Feature matching solves the problem of data association [1] in computer vision, which determine the correspondence between the features seen currently and those seen previously. By accurately matching the descriptors between images or between images and maps, a large burden can be reduced for subsequent posture estimation, optimization, and the like. However, due to the local characteristics of image features, mismatches widely exist and have not been effectively solved for a long time, which has become a major bottleneck restricting performance improvement in computer vision [2].

After feature points are detected in the process of image matching, feature descriptors are used to express the detected feature points in a certain mathematical way so that the machine can recognize them. Meanwhile, the uniqueness of the expression of feature

Y. Zhao et al. (Eds.): ICIG 2019, LNCS 11902, pp. 255–265, 2019.
https://doi.org/10.1007/978-3-030-34110-7_22

points is taken into consideration, so that mismatching will not occur. The ideal feature descriptor needs to satisfy the invariance of scale, rotation, and even affine transformation, and is not sensitive to noise. Only when the feature descriptors corresponding to different feature points have little correlation can different feature points be well distinguished. Therefore, improving the recognition and expression ability of feature descriptors is conducive to improving the overall matching quality. For image matching, local feature space distribution descriptor [3] is most widely used. One of the most representative studies is the SIFT descriptor. In 2004, KeY et al. developed the PCA-SIFT [4] descriptor by removing some insignificant direction gradient values through Principal Component Analysis (PCA), which significantly improved the speed. Mikolajczyk obtained the GLOH (Gradient Location and Orientation Histogram) descriptor [5] by using the expression of polar coordinate system instead of the expression expansion of European coordinate system. In 2008, Bay et al. proposed a 64-dimensional SURF descriptor to speed up the calculation. In 2010, Enign et al. proposed the DAISY [6] descriptor, which replaced the weighted calculation of some previous operators by means of convolution kernel, and it has a good application in the region with dense urban buildings. Local feature space association descriptors use a certain mathematical method to calculate the spatial correlation characteristics of local features such as gradient and binary [7]. In 2012, Vandergheynst et al. proposed a FREAK (Fast Retina Keypoint) descriptor based on the mechanism of human visual imaging [8], whose descriptors are more reasonable in terms of matching accuracy. In 2018, Yi et al. proposed LIFT [9] descriptor based on convolutional neural network, which can be used to learn descriptors and improve them compared with traditional manual descriptors.

Feature matching is the key to the rapid development of indoor positioning and navigation technology while traditional feature detection algorithms have different performance in different environments. Because indoor images are affected by illumination, angle and scale, a single traditional algorithm is not sufficient to meet the requirements in terms of processing efficiency and large rotation changes. Therefore, we performed feature matching on indoor images from the characteristics of image data, and compared processing speed and matching rate of the three traditional algorithms that of three methods using the improved triple-sample shallow convolutional neural network learning descriptors. Then we proposed an improved algorithm adopting the SURF feature detection, and the triple-sample shallow convolutional neural network learning descriptor, and we verified the performance of the algorithm under rotation changes using three groups of indoor image data in terms of matching rate, repetition rate and correct matching number.

2 Research Foundation

(1) SIFT Algorithm
The SIFT algorithm first constructs a Gaussian scale space through a Gaussian convolution kernel [10], and then performs extreme point detection and extraction on different scale space layers. In the SIFT detection method, after the spatial scale layer is constructed, the stabilized feature points are detected in the scale space by the function DOG (Difference of Gaussian). It is ensured that the SIFT

Algorithm has certain antagonistic effect on the scale change through the detection and calculation of feature points in the scale space. In the research of Prof. David Lowe, in order to reduce the impact of mutations, Gaussian function is also used to smooth the histogram.

(2) SURF Algorithm
The algorithm has been accelerated on the basis of SIFT, which makes it faster and more comprehensive. The SURF algorithm performs Gaussian filtering processing by adding and subtracting the integral image to speed up the construction of the scale space. The integrated of the image can be calculated by simply scanning the pixels on the original image. Simplify the Gaussian second-order differential template and perform a Gaussian convolution operation between the template and the image to convert it into a box filtering operation [11]. When constructing the image pyramid, the size of the box filter template is continuously expanded to obtain a linear scale space. The integral images and different sizes of filter templates are used to generate a response image of the Hessian matrix determinant, using a non-maximum suppression method. The feature point results in different scale spaces are obtained [12]. In order to make the feature points own rotational invariant performance, the Haar wavelet response calculation is used to determine the main direction of the feature points.

(3) ORB Algorithm
The ORB algorithm is currently the most widely used method in real-time image detection matching in the field of computer vision. It uses the OFAST algorithm to quickly perform feature point detection. The basic idea can be divided into two parts: the first part is the FAST corner extraction of the image, and seek the center point of the gray level obviously change; the second part is the construction of the BRIEF descriptor. A directional description of the surrounding area of the extracted feature points is made for subsequent matching of the feature points. The FAST involved in the ORB is a fast corner detection method, which mainly detects the position of the gray range change of the local range, compares the difference between the gray level of the central pixel and the gray level of the surrounding pixels, and determines the potential feature point when the difference is large. However, the detected feature points do not have scale and directionality. Therefore, in the ORB method, by constructing the image pyramid layer and detecting the corner points in each layer of the image pyramid, it is resistant to the scale change. The resistance to the rotation change is realized by the gray centroid theory, and the vector between the geometric center of the image and the center of the gray scale is calculated [13], giving the detection angle a main direction, causing the rotation of the image to change. Have a certain detection ability.

3 Improved Matching Algorithm Based on Neural Network Descriptor

The feature points extracted by the feature detection algorithm find the correspondence between images through local descriptors, which is one of the most widely studied problems in computer vision. Based on the end-to-end learning descriptor of CNN architecture [14], in the training of large data sets of positive and negative sample pairs, the core is to select the appropriate indoor image triple sample data, iterative optimize and update the network parameters through the principle of backpropagation algorithm to build good learning descriptors.

This paper improved the image feature matching based on the local feature descriptor TFEAT [15] based on the triple-sample shallow convolutional neural network learning. TFEAT utilizes training samples based on triple samples, as well as mining related information for difficult unrelated samples of triple samples. Difficult unrelated samples refer to the fact that the uncorrelated samples have relatively small values for the input network calculation output and are difficult to distinguish. In the image data, different objects in the room are selected to correspond to unrelated image blocks, and the parameters can be trained in negative. For example, due to the geometric transformation of image blocks corresponding to different objects, such as certain rotations, scales, etc., it is possible to make their performance in optical images consistent. Image blocks corresponding to the same object are very likely to be completely inconsistent in optical imaging. It is precisely because of the existence of such image blocks that the data set of "difficult negative sample pairs" needs to be fully utilized, and back propagation promotes shallow neural networks. The training makes the learning descriptors perform better.

Training using triple-samples involves samples the form of (a, p, n), where a (anchor) is the reference sample, p (positive) is the relevant sample, and n (negative) is the unrelated sample. In the training samples used in this dissertation, sample a and sample p are samples of different perspectives of the same feature point, and n is a sample of different feature points. Sample a and ample p in the feature space will close when optimizing network parameters, and pushes a and n away. See as Fig. 1.

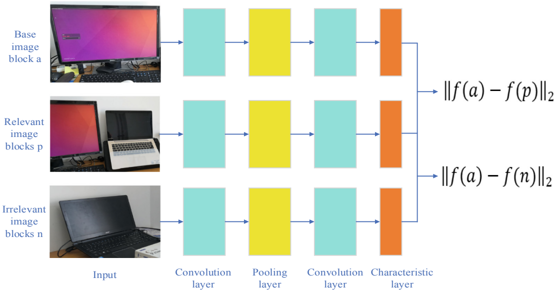


Fig. 1. Schematic diagram of the process of triple sample learning

The learning process of the above three samples can be expressed mathematically as formula (1):

$$\begin{cases} \delta_{+} = \| f(a) - f(p) \|_2 \\ \delta_{-} = \| f(a) - f(n) \|_2 \end{cases} \tag{1}$$

Where δ_{+} and δ_{-} represent similarities between sample features, the minimum value of the loss function is set to 0, there is no upper limit, and the network parameters are optimized to make the distance between a and p tiny. μ is a given parameter. When $\delta_{-} > \delta_{+} + \mu$, the value of the loss function drops to 0, the network parameters are no longer updated, the loss function of the learning training is formula (2), and the difficult unrelated samples in the triple sample are defined as formula (3):

$$\lambda(\delta_{+}, \delta_{-}) = max(0, \mu + \delta_{+} - \delta_{-}) \tag{2}$$

$$\begin{cases} \delta_{*} = min\left(\delta_{-}, \delta_{-}^{'}\right) \\ \delta_{-}^{'} = \| f(p) - f(n) \|_2 \end{cases} \tag{3}$$

When $\delta_{*} = \delta_{-}$, exchange the reference sample a and the related sample p in the calculation of the triple sample formula to make the related sample become the reference sample, and the reference sample becomes the relevant sample, which can make the indistinguishable irrelevance in the triple sample The sample is used for backpropagation. Through such a calculation process, the δ_{*} value will always be a set of sample pairs with a large feature similarity distance, and the mathematical calculation formula of the loss function at this time can be expressed as the formula (4).

$$\lambda(\delta_{+}, \delta_{*}) = max(0, \mu + \delta_{+} + \delta_{*}) \tag{4}$$

The learning descriptor effectively mines the difficult unrelated sample pairs to conduct network training, which reduces the total sample input and reduces the computational cost. For the initial time of network training, the network parameters are usually set to a random smaller number to achieve the purpose of initialization. The triple-samples shallow convolution training method was adopted to set the detailed parameters. After each iteration, the learning rate is updated to 0.9 times, the learning rate is set to 0.01, and iteration is continued until the drop is 10^{-6} to stop training. Due to the characteristics of the image data, after using the feature point matching algorithm to detect and match the image, there will be a certain mismatching point, which affects the accuracy of the subsequent data processing. Therefore, it is necessary to perform certain processing on the mismatched point. In this paper, the random sampling consistency algorithm (ie RANSAC algorithm) is adopted to eliminate the mismatched points.

Through experiments, the performance of the matching of the three traditional manual operators SIFT, SURF, ORB mentioned above and the performance matching based on neural network descriptors were verified by time. Two similar images were selected, and the image size was 4608 * 3456. In order to increase the running speed, it was sampled 4 times down to obtain an image of 1154 * 564. The traditional operator is based on windows7 (CPU: i3, graphics card GT-520M, 2G memory) and opencv2.4.9

programming environment implementation code, and the TFEAT descriptor matching was based on the Windows7's pytorch1.0.1 and opencv2.4.9 environment.

Since the shallow convolutional neural network learning of the triple samples in this paper is an improvement on the feature descriptor, the above three algorithms and methods combining the improved descriptor respectively are used to carry out experiments separately to verify the effectiveness of the algorithms with the improved descriptor. The number of feature points extracted by each algorithm was controlled to be about 1000, and the feature points of left and right images were preliminarily matched, then the RANSAC method was used to eliminate the mismatch, and the number of matching points is counted. The time and matching rate of each algorithm for detecting the same number of feature points were calculated and analyzed respectively, which are shown in Table 1.

Table 1. Feature point detection results

Detection operator	Points extracted separately from left and right images		The correct match points	Match time	Match rate
SIFT	923	745	351	6.05 s	42.1%
SURF	1045	1043	430	1.696 s	41.2%
ORB	1000	1000	287	0.878 s	28.7%
SIFT+TFEAT	1000	1000	456	4.758 s	45.6%
SURF+TFEAT	986	965	541	1.485 s	55.5%
ORB+TFEAT	752	718	258	0.644 s	35.1%

It can be seen from Table 1 that the algorithms using the three traditional algorithms for feature detection and the local feature descriptors based on the triple-sample shallow convolutional neural network learning in this paper has a greater processing efficiency and accuracy than the traditional algorithms. It can be seen from the analysis that the matching time through the SIFT operator combining with the TFEAT method to detect the same number of feature points reduced by 1.292 s, and the correct matching rate was increased by 3.5%. The traditional SURF combing with the TFEAT method has the highest correct matching rate, which reached 55.5%, and the matching time also reached by 0.21 s, which showed its obvious relative advantage. The least time-consuming is the ORB operator. Although the operation efficiency is highest but the matching rate is lowest, the matching accuracy is increased by 6.4% and the running time is reduced by 0.234 s after combining with feature descriptor of this paper for matching. Therefore, it can be verified that adopting the descriptor algorithm proposed in this paper based on the traditional feature extraction algorithm can effectively improve the accuracy of image matching and reduce the running time.

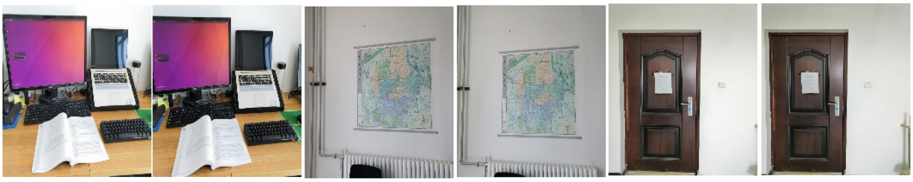
In summary, considering the real-time and matching rate, this paper selected the 64 * 64 image block with the feature points extracted by the SURF method in the indoor images as the center, and input the network for parameter calculation, then we obtained a 128-dimensional local learning descriptor after processing through the shallow convolutional neural network layer.

4 Experiment

The following experiment is mainly to verify the effectiveness of the improved algorithm combining the TFEAT descriptor based on the triple-sample shallow neural network for feature matching with the SURF algorithm for feature detection in terms of rotation changes. Aiming at the images with different features, we evaluated the performance of different algorithms including SIFT, SURF, ORB and our improved method in terms of rotation changes respectively, which chose feature point repetition rate and image matching accuracy as evaluation criteria.

4.1 Experimental Data Introduction

In this paper, three different types of indoor image data were used to analyze the performance of four different matching methods in three groups of data, and to compare and analyze the effectiveness of our proposed improved algorithm. The experimental data of the three groups were from different experimental areas in the room as shown in Fig. 2. The region a data were indoor desks, the data feature points were many, but a large number of textures were repeated. The region b data were the wall maps, the data features more but the features were not obvious. The region b data were the door frames, the number of data features were small, and there were a large number of areas such as white walls and almost no texture information.



region a region b region c

Fig. 2. Experimental image data

4.2 Feature Detection Algorithm Matching Experiment and Analysis

In this paper, four matching methods based on SIFT, SURF, ORB and our improved algorithm were used to carry out feature detection matching experiments in three types of images, and feature point repetition rate, correct matching quantity and image matching accuracy were used as evaluation criteria for performance evaluation and analysis.

(1) Repetition rate: refers to a rate obtained by dividing the number of feature points that can be repeatedly detected by the total number of detected features in an image pair with overlapping degrees. It can be obtained from the definition of repetition rate, which can reflect the adaptability of the detection algorithm on the image to some extent. The mathematical formula for the repetition rate is shown in (5):

$$R = \frac{N}{min(N_1, N_2)} h_{overlap} \tag{5}$$

Where R is the repetition rate, N is the number of feature points from the left image projection transformation to the right image, N_1 and N_2 are the feature points extracted from the left image and the right image respectively, and $h_{overlap}$ is the overlap degree of the left and right images.

(2) Matching accuracy rate: refers to the rate of the number of features matching correctly in the two images to the total number of feature matches. The correct matching number in this paper refers to the matching result of the two images under the homography transformation, and the corresponding difference of the same name image points is less than 1.5 pixels. Therefore, the matching correct number and the matching correct rate also reflect the matching accuracy and accuracy of feature points. The mathematical formula for the accuracy rate is shown in (6):

$$P = \frac{T}{N} \tag{6}$$

Where P is the image matching accuracy rate, T is the correct number of image matching, and N is the total number of image matching.

We respectively using four matching methods to detect the feature matching of each image relative to the reference image. In order to verify the performance of our proposed improved algorithm in the case of rotation, we need to set the data of 3 groups from 0° to 90 in 10° steps for image rotation and then feature extraction and matching, resulting in a large number of data experimental results, limited by the length of the article, we will only show statistical results. Detailed statistical analysis was performed on the number of feature points detected and matched for the above rotation changes. The comparison results of the repetition rate detection of the three groups data were shown in Fig. 3 (the abscissa is the magnitude of the rotation angle, and the ordinate represents the feature point repetition rate). The image matching accuracy was shown in Fig. 4 (the abscissa represents the rotation step and the ordinate represents the number of correct image matching), and the image matching accuracy was shown in Fig. 5 (the abscissa represents the rotation angle and the ordinate represents the correct matching rate). The time it takes for image matching is shown in Fig. 6 (the abscissa represents the rotation angle and the ordinate represents the time it takes to match). In all the graphs in this chapter, region a is represented on the left, region b is represented in the middle and region c is represented on the right.

When the image rotation changes, the feature points extracted by each algorithm were controlled at about 1000. As shown in Fig. 3, the repetition rate of the feature points extracted by each algorithm in region a -with obvious feature is relatively stable, whose

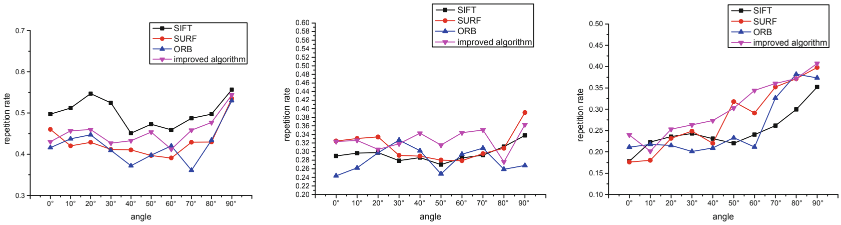


Fig. 3. Repeat rate comparison analysis

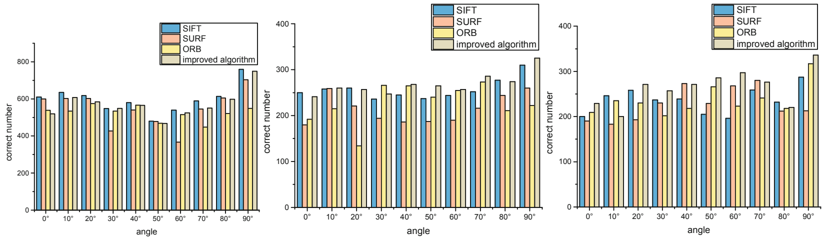


Fig. 4. Correct match quantity comparison

fluctuation range of the repetition rate is about 40%. In region b, the overall repetition rate of the improved algorithm is about 32%. In region c, the repetition rate fluctuation range of all algorithms is relatively large, and it can be seen that the improved algorithm is superior to the other three traditional algorithms. As shown in Fig. 4, the number of correct matches extracted by SIFT, SURF and our improved algorithm respectively is significantly higher than the ORB algorithm. It can be shown that although the number of repeated feature points extracted by the ORB can reach the number of above three algorithms, the number of correct extractions in the texture-like regions is significantly less than the previous two algorithms. In region c, it can be seen that the matching number of the improved algorithm is significantly higher than that of the three traditional methods, and the antagonistic effect on rotation is better. Figure 5 is a comparison of the matching rate of various algorithms. It can be seen that the matching rate of all algorithms in region a fluctuate significantly in the region a, and the improved algorithm performs best when the angles are rotated by 40° and 80°. In region b, it can be seen that our improved algorithm is weaker than the SURF algorithm when the angle is rotated

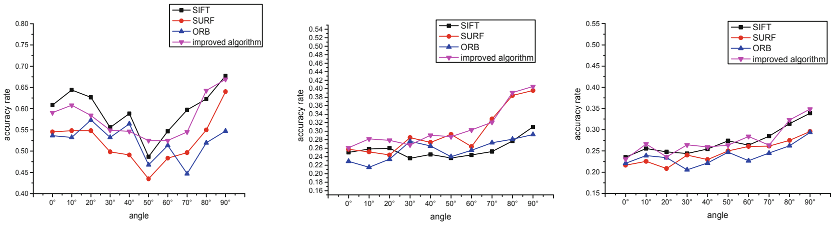


Fig. 5. Matching rate comparison analysis

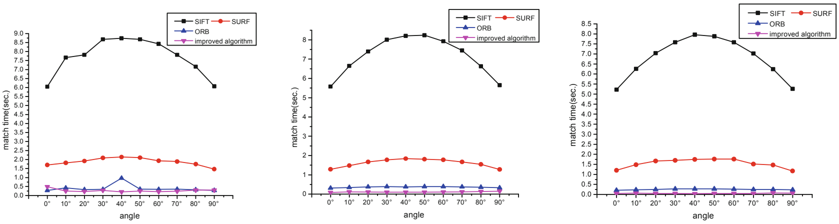


Fig. 6. Matching time comparison analysis

30° and the matching rate is higher than the other three traditional algorithms in the remaining rotation angles. In region c, when the image rotates 10°, 30°, 40°, 80° and 90°, the correct matching rate is significantly improved. Figure 6 shows the time required by the four algorithms when extracting the same feature points. It is obvious that the improved algorithm in this paper has obvious advantages.

In summary, the improved algorithm in this paper has a significant improvement in repetition rate, correct matching number and matching rate compared with other three traditional algorithms, especially in the environment where the feature points are sparse and the texture is weak. Rotational changes have certain antagonistic properties and can guarantee a certain reliability even in the case of large rotation angles.

5 Conclusion

Aiming at the problem that the traditional matching methods are difficult to balance the robustness and real-time of indoor image matching, this paper conducted image matching through three traditional algorithms and the three algorithms with the feature learning descriptor of the triple-sample shallow convolutional neural network respectively. It is verified by experimental comparison that the improved descriptor proposed in this paper can effectively improve the correct matching rate of images and improve the operating efficiency. Considering the requirements of real-time and accuracy of registration results comprehensively, we carried out image matching using SURF for feature detection and the improved feature learning descriptor of the triple-sample shallow convolutional neural network for feature matching, and used the appropriate data sample training method to obtain a better adapted model. We mainly verified the robustness of the algorithm in the case of rotation changes in terms of repetition rate, correct number and correct rate. The experimental results show that the improved algorithm has a smaller increase in feature-rich regions and a significant increase in regions without rich features. In addition, our improved algorithm has a stable fluctuation range, has certain antagonistic effect to the rotation changes, and has a good effect in regions with sparse features. Due to the limited amount of data and data range adopted in this paper, the next step is to use a variety of data to achieve image registration and environmental reconstruction for large scene indoor environments.

References

1. Chai, H.: Data association method for mobile robots in SLAM. Doctor, Dalian University of Technology (2010)
2. Gao, X., Zhang, T., Liu, Y.: Visual SLAM Fourteen Lectures –Form Theory to Practice, 1st edn. China Machine Press, Beijing (2017)
3. Yan, K., Sukthankar, R.: PCA-SIFT: a more distinctive representation for local image descriptors. IEEE Comput. Soc. **2**(2), 506–513 (2004)
4. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. In: Computer Vision and Pattern Recognition, USA, pp. 257–263 (2003)
5. Tola, E., Lepetit, V., Fua, P.: DAISY: an efficient dense descriptor applied to wide-baseline stereo. IEEE Trans. Software Eng. **32**(5), 815–830 (2010)
6. Geng, Z., Zhang, B., Fan, D.: Digital Photogrammetry. Surveying and Mapping Publishing House, Beijing (2010)
7. Leutenegger, S., Chli, M., Siegwart, R.Y.: BRISK: binary robust invariant scalable keypoints. In: IEEE International Conference on Computer Vision, Spain, pp. 2548–2555 (2011)
8. Vandergheynst, P., Ortiz, R., Alahi, A.: FREAK: fast retina keypoint. In: IEEE Conference on Computer Vision & Pattern Recognition, USA, pp. 510–571 (2012)
9. Yi, K.M., Trulls, E., Lepetit, V., Fua, P.: LIFT: learned invariant feature transform. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9910, pp. 467–483. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46466-4_28
10. Morevec, H.P.: Towards automatic visual obstacle avoidance. In: International Joint Conference on Artificial Intelligence, p. 584 (1977)
11. Simard, P.Y., Haffner, P., Lecun, Y.: Boxlets: a fast convolution algorithm for signal processing and neural networks. In: Conference on Advances in Neural Information Processing Systems, USA, pp. 571–577 (1999)
12. Hu, J.: Research on high-resolution aerial image feature matching technology. Doctor, East China University of Technology (2018)
13. Rosin, B.P.L.: Measuring corner properties. Comput. Vis. Image Underst. **73**(2), 291–307 (1999)
14. Fischer, P., Dosovitskiy, A., Brox, T.: Descriptor matching with convolutional neural networks: a comparison to SIFT. Comput. Sci. (2014)
15. Vijay, K.B.G., Carneiro, G., Reid, I.: Learning local image descriptors with deep siamese and triplet convolutional networks by minimizing global loss functions. In: IEEE Conference on Computer Vision & Pattern Recognition, USA, pp. 5385–5394 (2016)