A Review of Image Feature Descriptors in Visual Positioning

Wen Liu, Shuo Wang, Zhongliang Deng, (Senior Member, IEEE) Tingting Fan, Mingjie Jia, Hong Chen

School of Electronic Engineering, Beijing University of Posts and Telecommunications, Beijing, China

Abstract

Visual positioning is one of the important research directions of indoor positioning algorithms and systems. Image feature descriptor plays a key role in most visual positioning algorithms, which directly affects the speed and accuracy of positioning. This paper focuses on the application of image feature descriptors in visual positioning and studies image feature detection and extraction algorithms. The image feature descriptors are divided into three categories according to the characteristics of different, namely local gradient-based descriptors, image intensity-based descriptors, and learning-based descriptors. Which steps, characteristics, applicable scene, and application in visual positioning or image matching of the methods are studied. The main purpose of this paper is to make a review of the image feature descriptors that may be used in visual positioning, and provide a reference for the research and innovation of visual indoor positioning using image feature descriptors.

Keywords

Image features; feature descriptors, visual positioning

1. Introduction

The demand for location information is gradually increasing in daily life [1]. With the progress of science and technology, location-based services provide plenty of convenience for human life. Indoor robots gradually enter into life and provide services for human beings on various occasions. The location information of robots is the basis of these convenient services. Due to the serious blocking of satellite signals indoors, especially in buildings and underground space, it is impossible to obtain accurate indoor position information only relying on GNSS (Global Navigation Satellite System) [2]. Therefore, the research of indoor positioning is of great significance for providing location-based services when satellite signals cannot be relied on. For indoor service robots, quickly and accurately get current location information also provides a guarantee for their better services [3].

In the research of indoor visual positioning, the method using image features has achieved good results. The location, shape, and color of some objects change less indoors, which provides a good foundation for indoor positioning based on image features [4]. An important tool for the indoor location using image features is the image feature descriptor. It is a kind of simple and stable way to express visual information after the extracting of an image. Compared with the original image, it eliminates the redundant visual information which contributes less to image recognition and matching, and only focuses on the part with distinct "features", which are the parts that do not change easily with time and environmental conditions [5]. Good image feature descriptors used in indoor locations based on image features should be robust. They should maintain availability after environmental conditions, scale, or rotation changes. They should also have fast extraction and matching speed to meet the real-time requirements of indoor positioning [6]. Although some studies do not use feature points [7][8], which also achieve good results, this is not within the scope of this paper, so we will not explain them in detail in this review.

Proceedings International Conference on Indoor Positioning and Indoor Navigation, 29 Nov. – 2 Dec. 2021, Lloret de Mar, Spain
EMAIL: liuwen@bupt.edu.cn (Wen Liu); buptws@bupt.edu.cn (Shuo Wang); ftt 77@ bupt.edu.cn (Tingting Fan)
ORCID: 0000-0002-6450-1969 (Wen Liu);

© 2021 Copyright for this paper by its authors.
Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-3097/paper19.pdf

Image feature descriptors have been widely studied and applied in many fields, such as image matching, image retrieval, image classification, object recognition, target tracking, change detection, and so on [9]. In this paper, the traditional image feature descriptors are divided into two types: image feature descriptors based on local gradients and feature descriptors based on image intensity. Image feature descriptor based on local gradient refers to the feature descriptor calculated according to the image gradient information of the area around the image feature point. The image gradient information is saved as the feature descriptor of the interesting point. Euclidean distance is often used in subsequent matching, such as SIFT (Scale-Invariant Feature Transform), HOG (Histogram of Oriented Gradient), SURF (Speeded Up Robust). The feature descriptor based on image intensity refers to the feature descriptor calculated according to the relative size of the image intensity information of the pixel points or pixel blocks around the image feature points. The pixel intensity relationship is saved in the form of a binary sequence as the feature descriptor of the feature point. Hamming distance is used to measure the similarity in matching, such as BRIEF (Binary Robust Independent Element Features), ORB (Oriented FAST and Rotated BRIEF), BRISK (Binary Robot in Variant Scalable Keypoints), etc. In addition, the research on image feature descriptors extracted by machine learning has also made some achievements in recent years [10].

Compared with other reviews such as [11], this survey not only reviews the descriptors but also summarizes the application of various descriptors in localization. This paper makes notable contributions in the following aspects: 1) This review studies the commonly used image feature extraction methods. 2) This review provides a comprehensive overview of image feature descriptors and gives the classification of image feature descriptors, mainly discusses the classical descriptors and some excellent improvements. 3) In this paper, the research methods and achievements of various image feature descriptors in indoor visual positioning are described and combed. This paper analyzes and summarizes the image feature descriptors, aiming to provide ideas and improvement direction for the visual positioning method based on image feature descriptors and provide a reference for the research and innovation of indoor visual positioning in the future.

The following part of this paper summarizes the extraction methods and various descriptors according to the classification of image feature descriptors. The application of various descriptors will be summarized and the indoor positioning method based on image feature descriptors outlook will be discussed.

2. Extraction Methods of Image Feature Points

The generation process of image feature descriptors includes two stages: the extraction of image feature points and the description of image feature points [12]. Feature extraction refers to the process of extracting the points which can represent the image or some parts of the image. [13]. In general, feature points are represented by their position coordinates in the image. In this chapter, the basis, extraction method, and algorithm steps of image feature points selection in common image feature extraction algorithms are briefly introduced to help further understand the role and advantages of image feature points.

The image feature point extraction algorithm is the method to find robust and stable feature points in the image [14]. There are many methods for feature detection, which can extract the points or lines or other features through color or texture [15]. We mainly study the feature points used in indoor visual positioning. The feature points should be robust to the environment and will not lose stability due to the change of scale and rotation angle. The camera position or direction can still be determined according to these feature points when the environmental conditions change or the camera position and angle of view change.

Harris et al. proposed the Harris corner detection algorithm in 1988 [16]. The algorithm considers that most corners in the image can still show the characteristics of corners after changing the angle, so the corners can be used as image feature points. Figure 1 is the schematic diagram of the algorithm to judge the kind of zone. The algorithm calculates the gradient of pixels in two vertical directions, constructs the second-order response matrix in the form of a partial derivative, and gives the response function of the corner. Then it calculates the determinant and trace of the corresponding matrix of the corner and determines that the pixel is a corner, edge, or flat area according to the relationship between

the set coefficient and threshold value. Harris corner detection algorithm is a classic feature detection algorithm, which can detect corners stably, but has no scale invariance.

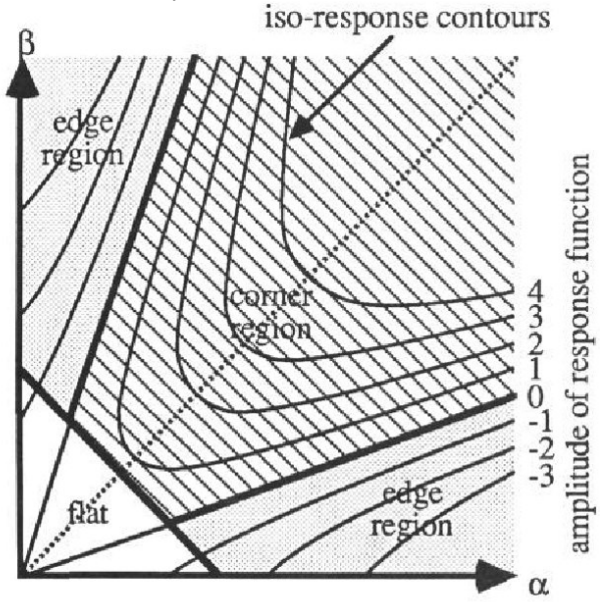

Figure 1: The schematic diagram of the Harris corner detection algorithm to judge the kind of zone[16]

In 1994, Jianbo Shi and Carlo Tomasi proposed a Shi-Tomasi corner detection algorithm based on the Harris corner detection algorithm [17]. It did not calculate the determinant and trace of the corner response matrix but directly compared the eigenvalue with the threshold value. If the smaller one of the two eigenvalues is greater than the minimum eigenvalue threshold, the strong corner will be obtained. Shi-Tomasi corner detection algorithm has the advantages of Harris corner detection algorithm and is faster than Harris corner detection algorithm.

Ojala et al. proposed a new image feature detection algorithm LBP (Local Binary Pattern) in 1996, which is used to extract local texture features of images [18]. LBP directly uses the information of image intensity to determine feature points, which provides a new idea for image feature point extraction and feature description.

SUSAN (Small Univalue Segment Assimilating Nucleus) was proposed by Smith et al. in 1997 [19]. Different from the fixed size square window used by Harris, the algorithm uses a circular window as is shown in Figure 2. Rather than sliding the window when judging whether a pixel is a feature point, it only compares the gray difference between the central pixel and other pixels in the circular area and counts the gray difference in the whole circular window. Because SUSAN compares the gray level of neighboring pixels in feature detection, it has a certain degree of rotation invariance. The selected circular template size has little effect on the selection of feature points, so it also has scale invariance to a certain extent.

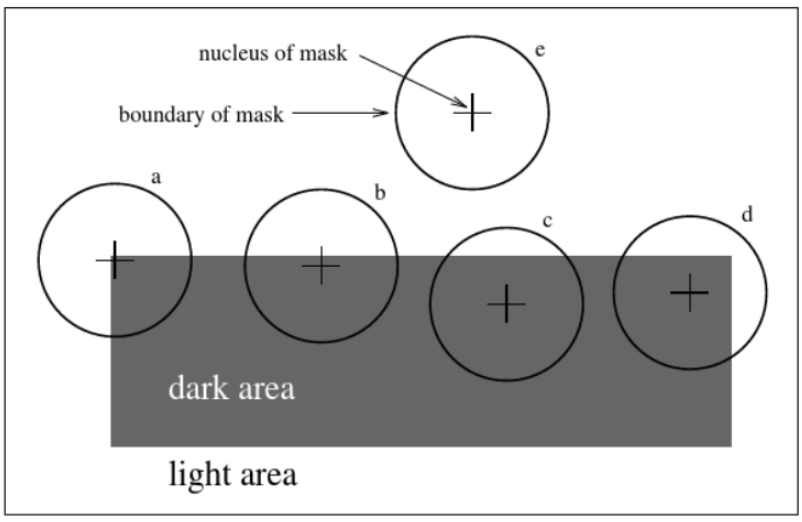

Figure 2: The schematic diagram of the circular window used in SUSAN[19]

In 1999, David Lowe downsampling the image constructed the scale pyramid, and convoluted it with the Gaussian kernel function to obtain the Gaussian difference pyramid. In the Gaussian difference pyramid, the extreme point was calculated as the feature point of the image [20]. Although the steps are complex, the DoG (Difference of Gaussian) descriptor is more robust and stable.

In 2006, Rosten et al. proposed FAST (Features From Accelerated Segment Test) [21]. We can get the basic principle of FAST in Figure 3. FAST detects the gray difference between 16 consecutive pixels and the center pixel in the 3 * 3 neighborhood around the point. If there are more than 12 consecutive pixels and the central pixel whose gray difference exceeds the threshold, the point is marked as the image feature point. The most obvious advantage of FAST is its speed. However, the image features detected by FAST have no scale invariance and rotation invariance.

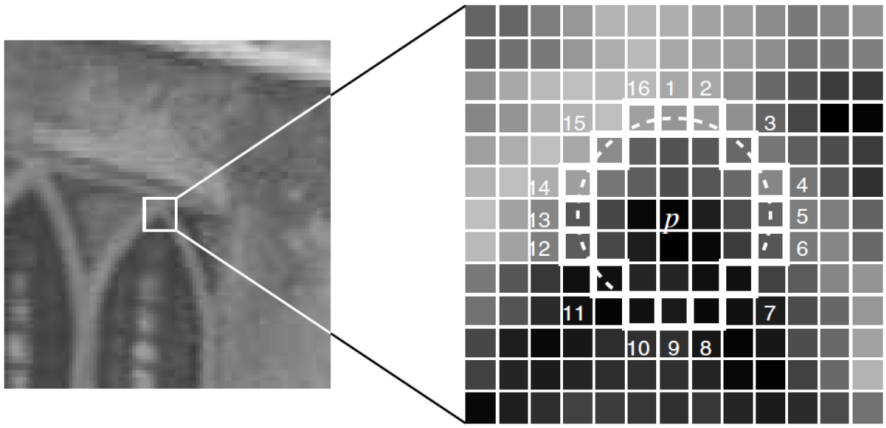

Figure 3: The Schematic diagram of sampling area when FAST feature detection algorithm extracts features[21]

Since most of the feature extraction algorithms and feature description algorithms are independent of each other, the feature description algorithm relying on image feature points can describe the feature points after the feature points are extracted by the above feature extraction algorithms. Therefore, how to select the image feature extraction method suitable for indoor visual positioning is the first step of the research work. In different application scenarios of indoor positioning, there are different ways to extract image features. If the positioning target is a moving object and needs to obtain the real-time location of the object to provide services, the weight of feature extraction algorithm efficiency should be increased when selecting feature extraction methods. Corner detection algorithm or FAST feature detection algorithm which are faster in the above algorithms should be considered. If the positioning target is relatively fixed or the surrounding environment doesn't change very often, and the real-time requirement is not high, a stronger robustness and stability algorithm such as the DoG feature detection algorithm should be chosen to obtain a better feature extraction effect. [22].

With the development of machine learning research, the extraction of these feature points can also be completed by machine learning. This paper also studies the image feature descriptors based on deep learning, which does not need some actual points in the image, so there is no actual "feature point extraction" process. The image feature descriptor based on deep learning will be discussed in detail in the following part of the next chapter.

3. Image Feature Descriptors

Feature description refers to the process of generating descriptions based on the extracted feature points according to the pixel gray level or image texture around the feature points. The feature descriptors are often stored in the form of vectors or matrices. We will focus on the influence of different description methods of image feature points on the matching performance of image feature descriptors in indoor visual positioning. According to the calculation method, feature descriptors are divided into three types: local gradient-based descriptors, image intensity-based descriptors, and learning-based descriptors.

3.1. Image Feature Descriptor Based on Local Gradient

Image feature description algorithm is applied to transform the extracted image information around feature points into low dimensional vector representation and store it. The vector representation obtained in this process is the image feature descriptor. Descriptors only pay attention to the part of information interested in the description algorithm and ignore the redundant information with low information entropy. At the same time, a low-dimensional vector or matrix can greatly save storage space.

The traditional image feature descriptors are divided into two types, one is based on local gradient and the other is based on image intensity. Image feature descriptor based on the local gradient is constructed by calculating and counting the histogram of gradient direction of the local area of an image. Local gradient image feature descriptor represented by SIFT (Scale Invariant Feature Transform) has been widely used in many image-based fields, such as image processing, computer vision, and so on.

SIFT is a landmark algorithm in the research of image local features. It was first proposed by David Lowe in 1999 and perfected in 2004 [23], which can detect and describe local features in images, find extreme points in scale space, and extract their position, scale, and rotation invariants. The core steps of SIFT feature extraction and description algorithm are divided into four steps, as showed in Figure 4. After two steps of Gaussian fuzzy processing, the image is input to scale space for extremum detection. The extremum points are mathematically processed to locate the key points. The image gradient method is used to determine the stable direction of the local structure. Finally, a group of vectors is used to describe image feature position, scale, and direction information.

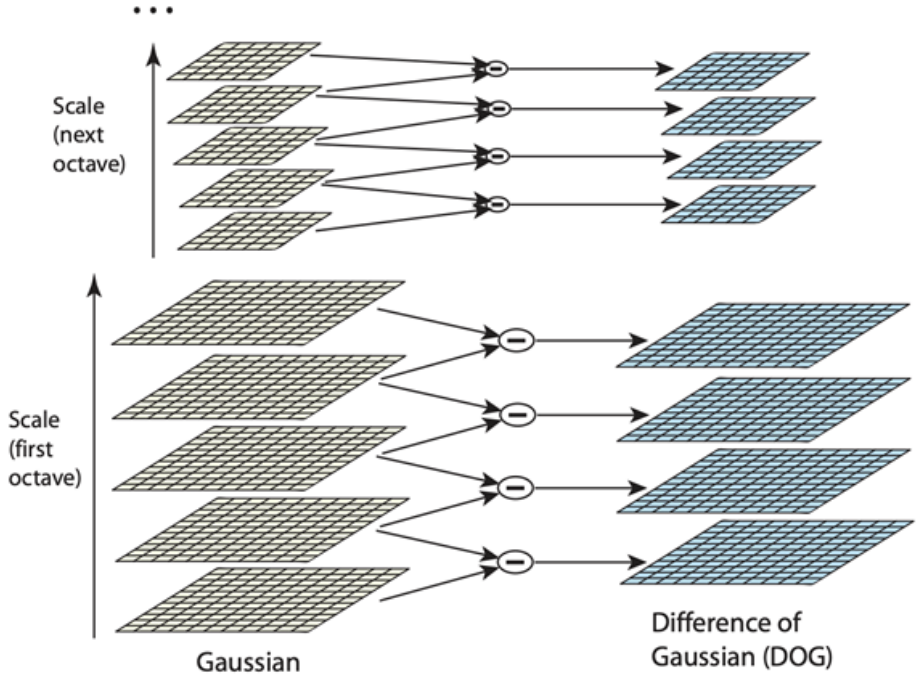

Figure 4: Gaussian scale space used in SIFT[23]

With the proposal of the algorithm, many subsequent scholars have carried out a variety of improvement and optimization designs around its core concept. Yan et al. improved the traditional SIFT algorithm in 2004 [24]. PCA (Principal Component Analysis) is applied to reduce the dimension of descriptors. The former steps of the algorithm (detecting feature points, modifying, calculating the main direction) are the same as those of SIFT. The difference lies in the construction of descriptors. Instead of using the original 128-dimensional descriptors, 39 * 39 * 2 gradient derivatives (horizontal gradient and vertical gradient of pixels in the neighborhood) are calculated on the 41 * 41 image block. Then PCA is used to get the eigenvector of the original 3042 dimensions to n (n is much less than 3042) dimensions. It can reduce the dimension of descriptors and filter out the interference information in some descriptors. At the same time, the algorithm opens a direction for future generations to study descriptors, such as GLOH (Gradient Location and Orientation Histogram). Mikolajczyk et al. proposed GLOH in 2005 [25]. It mainly improves the partition method around key points, from field grid partition

to eight quadrant circle grid partition, which enhances the robustness and independence of the descriptor. GLOH algorithm is based on log-polar coordinates when describing feature points. Three zones (6, 11, 15) and 8 angular directions are established in the radial direction in logarithmic polar coordinates. The detected region of interest is divided into 17 sub-regions according to the radial and angle. Then each region is divided into 16 bin gradients according to the SIFT descriptor method. The gradient angle is divided into 16 bins, and then the gradient histograms of each sub-region are spliced into a vector to get a 272-dimensional vector, namely 16 * 17. PCA is used to reduce the 272-dimensional vector to 40 dimensions.

SURF (speeded up robust features) is another improvement on SIFT proposed by Herbert et al. in 2008 [26]. Compared with SIFT, it is faster in execution. SURF uses the idea of simplified approximation in SIFT to simplify the Gauss second-order differential template in DOH so that only a few simple additions and subtraction operations are needed for the template to filter the image, and this operation is independent of the scale of the filter. Without down-sampling, the scale pyramid is constructed by keeping the image size unchanged but changing the size of the box filter, as showed in Figure 5. In the method of calculating the main direction of key points and the direction of pixels around the key points, SURF does not use histogram statistics but uses Haar wavelet transform. Relja et al. proposed RootSIFT [27] in 2012. Aiming at the similarity measurement method of SIFT features, a more accurate measurement method than Euclidean distance was proposed. After the description vector X of SIFT was extracted, the feature vector X was normalized by L1, and the square root of each element after normalization was calculated.

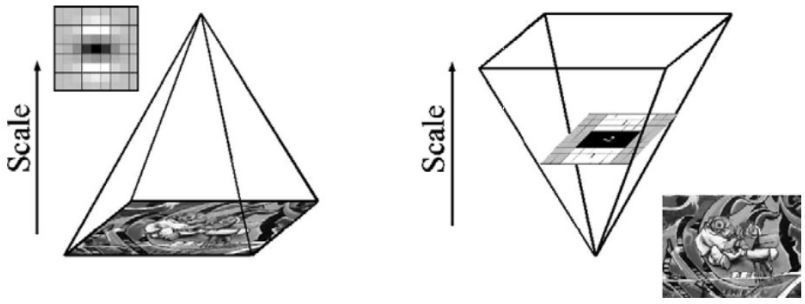

Figure 5: Instead of iteratively reducing the image size (left), the use of integral images allows the up-scaling of the filter at constant cost (right).used in SURF[26]

The image feature descriptor based on the local gradient is given in the form of a real vector after calculation, which can ensure scale, rotation invariance, and illumination robustness. However, due to the data form, the real-time performance of these descriptors is poor when they participate in the calculation (such as matching, clustering, etc.). How to improve the speed of participating in the operation while ensuring its advantages is the direction that needs to be paid attention to in the subsequent improvement of this kind of descriptors.

3.2. Feature Descriptor Based on Image Intensity

Another important descriptor is the feature descriptor based on image intensity. Since the LBP descriptors proposed by Ojala et al. in 1996, these descriptors have been improved and innovated and have been widely used. The basic idea is to express the image intensity difference in the local area around the feature points with binary string, then complete the image description and subsequent image matching in Hamming space [28].

LBP descriptors are often used to describe local texture features of images and are widely used in face recognition, facial expression recognition, and other fields. The basic LBP is defined in a 3 * 3 pixel neighborhood, the gray value of the center pixel is taken as the threshold value, as is shown in Figure 6. Then the gray value of the remaining 8 pixels in the neighborhood is compared with the threshold value, and the binary 1 and 0 are determined according to the comparison results. By connecting the eight binaries in order, we can get an 8-bit binary number, which is the LBP descriptor of a pixel. In 2002, Ojala et al. improved the LBP descriptors on scale and rotation invariance [29].

There are also many improvements on LBP feature descriptors, such as MB-LBP (Multi-Block LBP) [30], SEMB-LBP (Statistically Effective MB- LBP) [31]. LBP feature descriptors are characterized by low dimension, fast speed, rotation, and scale invariance.

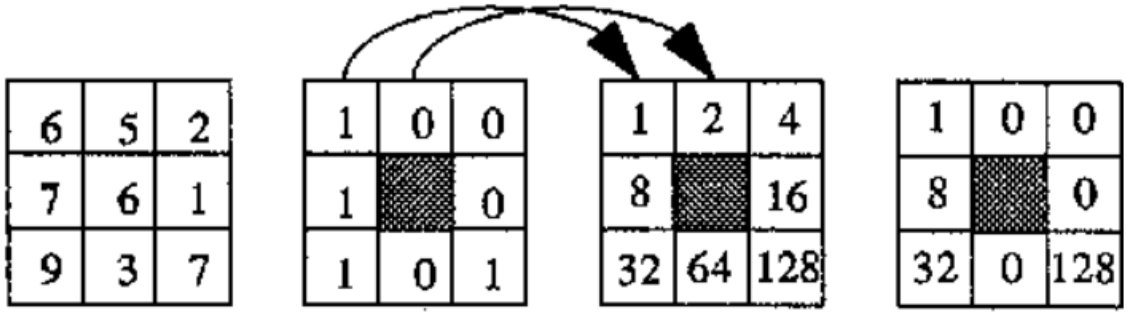

Figure 6: The schematic diagram of LBP algorithm when turning image intensity into binary descriptors in 3*3 pixel neighborhood [18]

In 2010, Calonder et al. proposed the BRIEF (Binary Robust Independent Elemental Features) feature description algorithm [32]. The strategy for selecting points in BRIEF is shown in Figure 7. When the before and after points follow a Gaussian distribution with the same parameters, the best robustness of the descriptor is obtained. Firstly, Gaussian filtering is applied to the image, and then two points are selected in the surrounding s * s pixel neighborhood with the feature points as the center. Then, the differences of gray values between the two points are compared, and binary assignment is performed. Finally, a set of 256-dimensional image feature descriptors is obtained. Although the BRIEF descriptor has the advantages in speed and robustness, it does not consider rotation in the description, and the matching performance is poor after image rotation.

Figure 7: Five point-taking methods used in the BRIEF Feature Description algorithm[32]. Respectively, Uniform Distribution, Co-Parametric Gaussian Distribution, Different Parametric Gaussian Distribution, Random Sampling and Polar Point Sampling. The Co-parametric Gaussian distribution is proved to be the most effective.

The emergence of the BRIEF descriptor further promotes the development of applications based on image feature descriptors. In 2011, Rublee et al. proposed ORB (Oriented FAST and Rotated BRIEF) feature detection and description algorithm on ICCV (IEEE International Conference on Computer Vision) [33]. In the aspect of feature extraction, ORB improves FAST to calculate the gray centroid of the image around the feature points and then records the vector from the center to the gray centroid. In the aspect of image feature description, rotation information is added to the description method compared with BRIEF. In the image feature description stage, the image blocks around the feature points are rotated according to the "main direction" of the features, and then the image is described by BRIEF. ORB selects FAST as the feature extraction method, and the improved BRIEF for description, which makes it have a great advantage in speed. The introduction of direction increases the robustness to rotation.

Also on ICCV in 2011, Leutenegger et al. proposed BRISK (Binary Robot Invariant Scalable Keypoints), which is robust to noise and has rotation invariance [34]. In the image feature extraction, BRISK also uses the method of main direction similar to ORB to ensure the rotation invariance. BRISK uses a uniform sampling pattern, which is shown in Figure 8. It is to construct concentric circles with a different radius centered on feature points and obtain a certain number of equal interval sampling points. Because this neighborhood sampling mode will cause overlapping effects, Gaussian filtering is needed for sampling points on concentric circles. The BRISK descriptor can achieve better matching performance when processing fuzzy images.

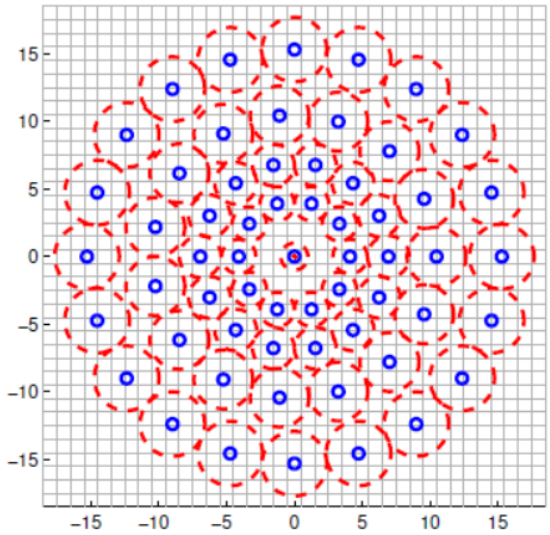

Figure 8: The schematic diagram of the concentric sampling pattern used in BRISK algorithm[34]

In 2012, Alahi et al. proposed FREAK (Fast Retina Keypoint) feature descriptor on CVPR (IEEE Conference on Computer Vision & Pattern Recognition) [35]. Based on the distribution of human retinal cells, They proposed a kind of sampling mode with the dense and sparse periphery, so that many overlapping sampling areas were constructed in the image. Compared with BRISK, the same method is used for concentric circles. All sampling points are located in concentric circles, as shown in Figure 9, but the difference is that sampling points are not evenly distributed, and the radius of the sampling area around sampling points is different. In the uniform sampling mode adopted by BRISK, the distribution of sampling points is uniform, and the size of sampling areas around sampling points is the same. It makes more sampling points on the outer concentric circle. In the sampling mode adopted by the FREAK descriptor, the number of sampling points on the concentric circle is the same (the original algorithm selects 6 sampling points on each circle). Therefore, the closer to the center, the more dense the sampling points are, the smaller the radius of the surrounding area (also referred to as "sense field"), and the more overlapping areas are. The distribution of sampling points in peripheral areas is relatively sparse. The outstanding part of FREAK is that the visual mechanism of the human retina is applied into random point pair sampling mode. The actual performance is not as good as ORB, but it is more robust to illumination.

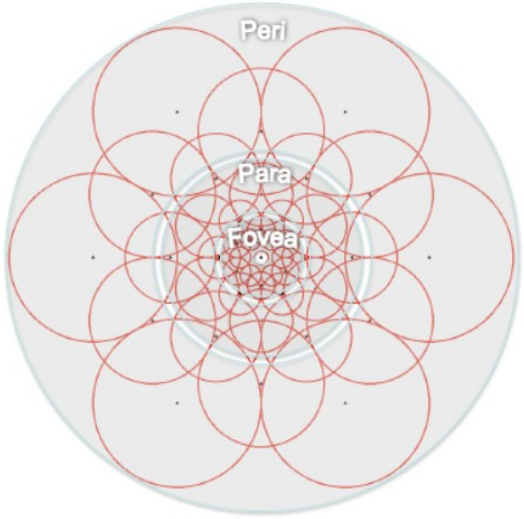

Figure 9: The schematic diagram of the sampling pattern used in FREAK algorithm[35]

The descriptors based on image intensity get binary descriptors after comparing the image intensity, which makes the descriptor based on image intensity have the advantage of matching speed from the structure of descriptors. However, due to the simplicity of the algorithm and the binary representation of descriptors is determined by the relationship of image intensity, it is greatly affected by lighting

conditions. Therefore, in addition to the rotation and scale invariance, how to make the descriptors more robust to illumination changes should be considered when improving and innovating such descriptors.

3.3. Feature Descriptor Based on Learning

Machine learning and deep learning have developed rapidly in recent years, and are widely used in image processing. Especially after 2010, the combination of machine learning and traditional descriptors and learning methods to improve the extraction performance have become one of the hot directions of descriptors in recent years. Abandoning the traditional image feature descriptors algorithm flow, and finding a new way, the method of directly processing the image through deep learning to get the visual information description opens up a new direction of image processing. This method does not detect the feature points alone, but adaptively finds the feature description of the image according to the task by gradient descent according to the network structure and loss function.

Among the networks used in deep learning, the typical one is CNN (Convolutional Neural Networks). CNN is a kind of feedforward neural network that contains convolution computation and deep structure. It is one of the representative algorithms of deep learning. CNN has the ability of representation learning, which can classify the input information according to its hierarchical structure. Based on the basic model of CNN, LeNet5 marks the real appearance of CNN [36]; AlexNet is the first model that introduces convolutional neural networks into the field of computer vision and achieves breakthrough results [37]. Visual Geometry Group of Oxford proposed VGGNet and won the ILSVRC 2014 competition champion with a 25.3% error rate [38]. The 16-layer VGG version is shown in the figure 10, The concatenation of multiple convolutional blocks can effectively extract image features while making the network deeper.

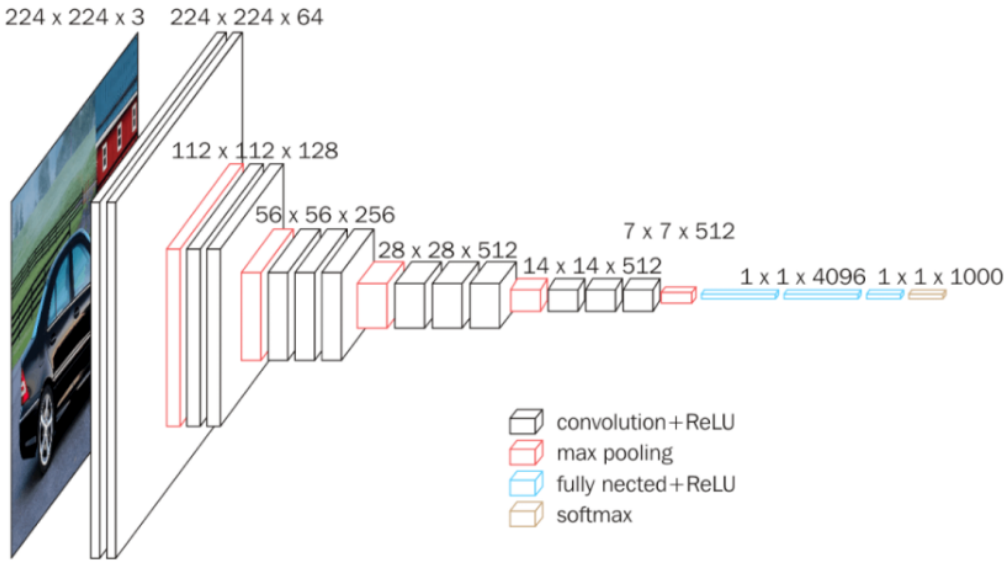

Figure 10: The schematic diagram of the sampling pattern used in FREAK algorithm[35]

Veit et al. of Microsoft Asia Research Institute proposed ResNet (Residual Neural Network) [39], successfully trained 152 layer deep neural network with the use of a residual unit, won the championship in ILSVRC2015 competition, the top-5 error rate is 3.57%, and the parameter quantity is much lower than VGGNet. Google's EfficientNet [40] further optimizes image recognition performance through compound scaling methd. And EfficientNet-B7 got the top-1 accuracy rate of 84.4% and the top-5 accuracy rate of 97.1% on the ImageNet data set. Compared with other models with the highest accuracy rate at that time, the parameter amount has been reduced by 8.4 times and the efficiency has been increased by 6.1 times. The basic models of CNN play an important role in the field of computer vision. On the one hand, a lot of work is based on these models as backbones. These models provide a mapping method from image to feature descriptor, which provides a better image representation for tasks.

Yi et al. proposed a new deep network structure in 2016 [41], which realizes the complete process of feature point processing, namely detection, direction estimation, and feature description. The three

components are formed into a separable network by using spatial converter and soft-argmax function. Effective training methods and difficult sample mining strategies are designed specifically for this network to improve the discrimination of samples. In addition to the combination of depth network and feature point processing, Tateno et al. proposed to use a neural network to predict the depth map to carry out accurate and dense monocular reconstruction in 2017, naturally fusing the dense depth map predicted by CNN with the depth data measured by direct monocular SLAM [42]. In the fusion scheme, the depth prediction has priority in the image location with the large error of monocular SLAM (such as low texture area), while the priority of depth prediction is reduced in the image location with complex texture, which relies on more sufficient image feature descriptors for monocular reconstruction. The introduction of a depth network improves the accuracy and repeatability of monocular reconstruction. In addition, the combination of deep learning and traditional descriptors can also achieve adaptive feature extraction [43]. In 2018, Detone et al. proposed a full convolution neural network structure for interest point detection and description. The structure introduced the homographic adaptation method (a multi-scale, multi-source method that can improve the stability of interest point detection), and transformed sparse interest point detection and description into an efficient adaptive convolution neural network. Also in 2018, Ruihao Li et al. Proposed a new monocular unsupervised deep learning system UnDeepVO [44]. The system uses space loss and time loss between stereo image sequences for unsupervised training. During the test, the system can estimate the attitude and dense depth map of monocular images. This system is different from other model-based or learning-based monocular virtual reality methods. The scale restored in the training phase improves the recovery accuracy. The VO method based on unsupervised learning has the potential to improve performance with the increase of training data set. In addition, many types of research combine deep learning with traditional descriptors. They fully combine the advantages of deep learning automatic extraction with a fine-grained description of traditional descriptors and provide new ideas for the description of feature points.

The learning-based feature descriptor can be understood as the features collected from each layer of the network. Compared with the traditional descriptors, this is variable and more adaptive. For CNN, the feature descriptor is extracted in the form of the sliding window by using the convolution kernel learned, and the feature descriptor of the image is obtained by some activation functions such as Relu. The difference of characteristics between different layers depends on the scope of the receptive field and also on the task. It was mentioned in [45] that in the object detection task, the color information is the most felt by the bottom receptive field, and the number of perception units for the object will increase significantly with the increase of the number of layers. In the scene recognition task, the bottom receptive field has a stronger perception of objects and some areas of objects. When the number of layers rises, it will focus on the color of the scene. Each layer in the network extracts a feature descriptor, and the latter extracts and integrates from the previous descriptors to get a better embedding, which gradually approaches the feature space of ground truth. Therefore, from the bottom layer of the network to the high level of the network, the extracted descriptor will be more suitable for tasks.

There are several innovative directions for researchers' reference. When extracting feature points, It can be chosen whether to use an image scale pyramid or not according to whether scale invariance is needed, as well as the feature direction for rotation invariance. In addition, the distribution of feature points should be more uniform to avoid the direct influence of partial feature disappearance and shadow caused by illumination condition change on image matching. The descriptor is a sparse image representation. The dimension of the image feature descriptor is related to storage space and matching speed. The feature dimension or the number of stored feature descriptors should be reduced under the demands of the matching performance. The detected feature points and descriptors can be integrated according to the matching performance to avoid repeated descriptions.

4. Application of descriptors

SIFT descriptors and various optimized descriptors derived from SIFT have been widely used in various image-based algorithmic scenarios because of their scale, rotation invariance, and the advantages of effectively suppressing the influence of illumination and noise. Image mosaic technology is needed in computer vision, virtual reality, remote sensing image processing, video monitoring, and other fields. The most important role of image mosaic is image feature descriptor. The quality and

efficiency of mosaic will be directly determined by the quality of feature extraction and description algorithm. SIFT algorithm is recognized as a more effective and stable method because of its advantages [46]. As one of the key technologies in digital image processing, image registration technology has become the core foundation of image mosaic [47] and object detection and tracking [48]. Image registration is the process of spatial matching of two or more images collected by different sensors under different environmental conditions. SIFT descriptors are often used in feature descriptions of image registration [49]. It is recognized as a more effective and stable method. The main conventional algorithm used in image retrieval and classification is BoW [50]. The core idea is to extract the key point descriptors and train a codebook by the clustering method. Then the number of times each descriptor vector in each image appears in the codebook is used to represent the image. SIFT descriptor is often used in the keypoint descriptors.

In the field of indoor visual positioning, the feature descriptors play an important role in many kinds of algorithm models. For example, the hierarchical structure position estimation algorithm [51] proposed by SMILE Lab of School of Electronics and Information Engineering, Xi'an Jiaotong University in 2013 used SIFT to represent the local features of the images in the dataset, and the SIFT obtained are used for subsequent clustering work. In 2015, the Multimedia Computing Group of Delft University of Technology proposed the Geo-Visual Ranking [52], which only uses visual images to predict geographic coordinates. The candidate image selection in the model relies on SIFT descriptor for image characterization and similarity measurement. In 2015, Deretey et al. Proposed a method for indoor positioning using a monocular camera [53]. In the process of generating a 3D map in the offline phase, the SIFT / SURF descriptors are used for 2D-3D feature projection.

Binary descriptors are widely used in face recognition, object recognition, SLAM (Simultaneous Localization And Mapping) scenes because of their advantages such as relatively simple calculation, low feature dimension, and fast matching speed. The system of positioning and navigation based on binary descriptors complements other data sources and achieves good results. Qian et al. Proposed a fast face recognition system based on PLBP (Pyramid of Local Binary Pattern) [54] and RI-LBP (Rotation Invariant Local Binary Pattern) [55]. Hussain et al. Developed an LQP (Local Quantified Pattern) technology for face recognition [56].

The advantages of fast speed, low dimension, and wide applicability of the binary descriptor also make it popular in the direction of visual positioning. Qin et al. developed a set of high-quality visual-inertial navigation positioning systems, namely VINS-MONO, and carried out open-source sharing [57]. VINS uses BRIEF as the descriptor of the system's visual positioning and loop detection. After the features are extracted by the FAST feature extraction algorithm, the BRIEF feature description algorithm is used to describe and save the obtained descriptor, and the sparse point cloud is constructed according to the monocular camera. Through the mapping relationship between the matching features of the two images, the motion of the robot is calculated when images are acquired. At the same time, the feature descriptor point cloud map is constructed to judge whether the robot has been to its current position. This kind of loop detection technology is very important to correct the positioning error. The loop detection makes good use of the advantages of the low dimension, easy storage, space-saving, and fast matching speed of BRIEF, which ensures the real-time positioning to a certain extent. ORB is also popular in applications with high requirements of real-time and stability, such as the ORB-SLAM system with better implementation performance in the SLAM field [58]. Based on the ORB feature detection and description algorithm, this system builds a set of an indoor positioning systems for real-time positioning and mapping. The positioning and pose optimization speed is fast, the ORB sparse point cloud map is stable, and the comprehensive accuracy, speed, and resource utilization are better than a dense map. Different from VINS, which uses an inertial navigation algorithm to assist positioning, the ORB-SLAM system's positioning process relies only on visual information, which makes it more difficult. VINS uses the data of the inertial measurement unit to calculate the rotation of the robot, and then uses the obtained rotation information for the extraction of visual information. Therefore, VINS only needs to use the BRIEF descriptor to achieve good results. ORB-SLAM is to introduce rotation information from the feature extraction stage and rotate the local image according to the rotation information in the description stage. It can achieve better results without assistance and makes the visual positioning results more robust to rotation. ORB-SLAM constructs the feature descriptor point cloud map for loop detection. The positioning error is corrected by matching the currently extracted feature descriptor with the descriptor stored in the database, to further improve the positioning accuracy. The

matching speed of ORB feature descriptors of the same dimension is the same as that of BRIEF, so the real-time positioning can be guaranteed. Leutenegger et.al. developed a visual-inertial navigation positioning system based on the BRISK descriptor in 2013, namely OKVIS [59]. This is a real-time positioning system using a stereo camera and IMU. After using the Harris corner detection algorithm for feature detection, BRISK is used for feature description. In positioning, the IMU is used to obtain the predicted value of the "next position". During feature matching, feature matching is performed between the current frame and the previous frame. Between the current frame and the image frame of the predicted position, the feature descriptors that can be matched are selected. The mapping relationship between the feature points in the 3D space and the image is established through the image matching results. In feature matching, instead of guiding the matching region before matching, OKVIS directly uses violence to find the descriptors that can complete the matching in the selected descriptors, which can also ensure the real-time positioning to a certain extent. However, OKVIS has no closed-loop detection or correction scheme, so it is easy to accumulate positioning drift after long-time or long-distance movement, which reduces the accuracy of positioning results. Moreover, in terms of CPU utilization, OKVIS using Harris corner detection algorithm and BRISK is higher than VINS-MONO which uses FAST and BRIEF.

The computer vision field based on the CNN algorithm derives many subdivision application scenarios, such as camera relocation, target detection, instance segmentation, scene recognition, depth estimation, image alignment, etc. There may be more subdivisions based on different tasks, such as vehicle re-identification, pedestrian detection, and other tasks in the instance segmentation field.

Indoor positioning involves camera relocation, image retrieval, and other fields. The camera relocation task includes the prediction of six degrees of freedom of the camera: the position and height of the camera, and the rotation angle of three dimensions. RelocNet proposes a convolutional network representation learning method for camera pose retrieval [60], which is based on feature descriptors of nearest neighbor matching and continuous metric learning. The feature embedding network is optimized by using the overlap information of camera cones between image pairs. The difference between the final camera pose descriptors of the network represents the change of camera pose. In addition, they also constructed a pose regression, which uses geometric loss to train, and obtain a more precise relative pose relationship between images based on query image and nearest neighbor image. Experimental results show that the proposed method can be generalized in different datasets. In 7Scenes and RelocDB datasets, the matching success rate is about 70%, which is better than other methods. Based on the image retrieval method, the two-dimensional plane position information of the camera is obtained by comparing the image with the fingerprint collected from different places. The input image can be a picture in the same direction, a panoramic picture, or a picture based on some specific objects. The image data processed by [61] collect images from four directions (front, back, left, and right). After the fingerprint map of the network is established, the input image will use the features extracted by the convolution network, and then input into Graph Location Networks for positioning. Finally, the matching accuracy reaches 93.92% in ICUBE.

Image feature descriptor is one of the key technologies for image matching, and it has been widely used in many fields. In the field of localization, a trade-off should be made between the speed of the image feature descriptor and the robustness of different conditions. The image feature descriptor which is more suitable for the task should be selected. The innovative image feature descriptor should be designed according to the needs. In addition, relying solely on image feature descriptors for indoor positioning has many difficulties to overcome, and the performance might be not so satisfactory. More and more researchers tend to use multi-sensor fusion for localization, which also provides a new direction and possibility for the expansion and application of image feature descriptors.

5. Conclusion

In this review, we conduct a comprehensive overview of the various image feature descriptors that may be used in visual positioning. For each kind of descriptor, we summarize the principle and improvement of the algorithm in detail. We provide a thorough review, comparisons, and summarizations of descriptors and their applications in recent years. Finally, we summarize the

applications of features descriptors in image matching or indoor visual positioning. We hope this paper can provide innovative ideas and references for researchers.

6. Acknowledgement

This work was financially supported by the National Natural Science Foundation of China (No.61871054).

7. References

[1] Dardari D, Closas P, Djurić P M T. Indoor tracking: Theory, methods, and technologies. IEEE Transactions on Vehicular Technology 2015, 64(4): 1263-1278.
[2] Pu Y C, You P C. Indoor positioning system based on BLE location fingerprinting with classification approach. Applied Mathematical Modelling 2018, 62: 654-663.
[3] DeSouza G N, Kak A C. Vision for mobile robot navigation: A survey. IEEE transactions on pattern analysis and machine intelligence 2002, 24(2): 237-267.
[4] Martín-Gorostiza E, García-Garrido M A, Pizarro D, et al. An indoor positioning approach based on fusion of cameras and infrared sensors. Sensors 2019, 19(11): 2519.
[5] Leng C , Zhang H , Li B , et al. Local feature descriptor for image matching: a survey. IEEE Access 2018:1-1.
[6] Qin Q , Luo B , Wu C , et al. A Local Feature Descriptor Based on Combination of Structure and Texture Information for Multispectral Image Matching. IEEE Geoscience and Remote Sensing Letters 2019.
[7] Demirhan M , Premachandra C . Development of an Automated Camera-Based Drone Landing System. IEEE Access 2020, 8:202111-202121.
[8] Premachandra C , Thanh D N H , Kimura T , et al. A study on hovering control of small aerial robot by sensing existing floor features. IEEE/CAA Journal of Automatica Sinica 2020, 7(4):1016-1025.
[9] Nunes C F G , Flávio L. C. Pádua. A local feature descriptor based on log-gabor filters for keypoint matching in multispectral images. IEEE Geoscience and Remote Sensing Letters 2017, PP(10):1-5.
[10] Zeng Z , Zhang J , Wang X , et al. Place recognition: an overview of vision perspective. Applied Sciences 2018, 8(11).
[11] Mikolajczyk K , Schmid C . A performance evaluation of local descriptors. IEEE Transactions on Pattern Analysis and Machine Intelligence 2005, 27(10):1615-1630.
[12] Li Q , Wang G , Liu J , et al. Robust scale-invariant feature matching for remote sensing image registration. IEEE Geoence & Remote Sensing Letters 2009, 6(2):287-291.
[13] Gauglitz S , Hllerer T , Turk M . Evaluation of interest point detectors and feature descriptors for visual tracking. International Journal of Computer Vision 2011, 94(3):335.
[14] S.Gauglitz, T. Hollerer, P. Krahwinkler. A setup for evaluating detectors and descriptors for visual tracking. IEEE International Symposium on Mixed & Augmented Reality. IEEE 2009.
[15] Premachandra C , Murakami M , Gohara R , et al. Improving landmark detection accuracy for self-localization through baseboard recognition. International journal of machine learning and cybernetics 2017, 8(6):1815-1826.
[16] Harris C G , Stephens M J . A combined corner and edge detector. Alvey vision conference 1988.
[17] Shi J , Tomasi C . Good features to track. Proceedings / CVPR, IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2002,600.
[18] Ojala T , Pietikainen M , Harwood D . Performance evaluation of texture measures with classification based on Kullback discrimination of distributions. Pattern Recognition, 1996. Vol. 1 - Conference A: Computer Vision & Image Processing. Proceedings of the 12th IAPR International Conference on 1996.
[19] Smith S M , Brady J M . SUSAN—A new approach to low level image processing. International Journal of Computer Vision 1997, 23(1):45-78.

[20] Lowe, D.G. Object recognition from local scale-invariant features. Proceedings of the International Conference on Computer Vision 1999, 1150-1157.
[21] Rosten E . Machine learning for high-speed corner detection. European Conference on Computer Vision. Springer-Verlag 2006.
[22] Khwildi R , Azza Ouled Zaid. HDR image retrieval by using color-based descriptor and tone mapping operator. The Visual Computer 2019, (8).
[23] Lowe D G . Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 2004, 60(2):91-110.
[24] Yan Ke, Rahul Sukthankar. PCA-SIFT: A more distinctive representation for local image descriptors. Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'04) 2004, 1063-6919/04.
[25] Krystian Mikolajczyk and Cordelia Schmid "A performance evaluation of local descriptors". IEEE Transactions on Pattern Analysis and Machine Intelligence 2005, 10, 27, pp 1615--1630.
[26] Herbert Bay , Andreas Ess, "SURF: Speeded up robust features". Computer Vision and Image Understanding (CVIU) 2008, Vol. 110, No. 3, pp. 346–359.
[27] Relja Arandjelović; Andrew Zisserman, Three things everyone should know to improve object retrieval. IEEE Conference on Computer Vision and Pattern Recognition 2012, 16-21 June.
[28] Yongqiang G , Yu Q , Weilin H . Local binary descriptors and its application to image matching. Journal of Network New Media 2014.
[29] Ojala T , Pietikainen M , Maenpaa T . Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE 2002:971-987.
[30] Liao S C , Zhu X X , Lei Z , et al. Learning multi-scale block local binary patterns for face recognition. In Lecture Notes in Computer Science 4642; Center for Biometrics and Security Research & National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, 95 Zhongguancun Donglu, Beijing 100080, China, 2007.
[31] Ebadi S , Ansari N N , Naghdi S , et al. The effect of continuous ultrasound on chronic non-specific low back pain: a single blind placebo-controlled randomized trial. Bmc Musculoskeletal Disorders 2012, 13.
[32] Calonder M , Lepetit V , Strecha C , et al. BRIEF: Binary robust independent elementary features. Computer Vision - ECCV 2010, 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV. Springer, Berlin, Heidelberg 2010.
[33] Rublee E , Rabaud V , Konolige K , et al. ORB: An efficient alternative to SIFT or SURF. International Conference on Computer Vision. IEEE 2012.
[34] Leutenegger S , Chli M , Siegwart R Y . BRISK: Binary robust invariant scalable keypoints. International Conference on Computer Vision. IEEE 2011.
[35] Alahi A , Ortiz R , Vandergheynst P . FREAK: Fast retina keypoint. IEEE Conference on Computer Vision & Pattern Recognition. IEEE 2012.
[36] El-Sawy A, Hazem E L B, Loey M. CNN for handwritten arabic digits recognition based on LeNet-5. International conference on advanced intelligent systems and informatics. Springer, Cham 2016: 566-575.
[37] Krizhevsky A , Sutskever I , Hinton G . ImageNet classification with deep convolutional neural networks. NIPS. Curran Associates Inc 2012.
[38] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint 2014:1409.1556.
[39] Veit A, Wilber M J, Belongie S. Residual networks behave like ensembles of relatively shallow networks. Advances in neural information processing systems 2016: 550-558.
[40] Tan M, Le Q. Efficientnet: Rethinking model scaling for convolutional neural networks[C]//International Conference on Machine Learning. PMLR, 2019: 6105-6114.
[41] Yi K M , Trulls E , Lepetit V , et al. LIFT: Learned invariant seature transform. European Conference on Computer Vision. Springer, Cham 2016.
[42] Tateno K , Tombari F , Laina I , et al. CNN-SLAM: Real-time dense monocular SLAM with learned depth prediction. Computer Vision and Pattern Recognition (CVPR). IEEE Computer Society 2017.
[43] D. Detone, T. Malisiewicz, and A. Rabinovich. Superpoint: Self-supervised interest point detection and description. CVPR Workshop on Deep Learning for Visual SLAM 2018.

[44] R. Li, S. Wang, Z. Long, and D. Gu, “UnDeepVO: Monocular visual odometry through unsupervised deep learning,”. arXiv preprint arXiv 2017:1709.06841.
[45] Bau D, Zhu J Y, Strobelt H, et al. Understanding the role of individual units in a deep neural network. Proceedings of the National Academy of Sciences 2020.
[46] LUO J,GWUN O. A comparison of SIFT, PCA-SIFT and SURF. Inernational J. Image Processing (LP) 2009, 3(4) :143-152.
[47] Zhang F,Liu F.Parallax-tolerant image stitching. 2014IEEE Conference on Computer Vision and Pattern Recognition(CVPR).IEEE 2014:3262-3269.
[48] Barbara Zitova, Flusser J.Image registration methods: a survey. Image and Vision Computing 2003,21(11):977-1000.
[49] Adel E,Elmogy M,Elbakry H. Image stitching based on feature extraction techniques:a survey. International Journal of Computer Applications 2014, 99(6):1-8.
[50] Hervé Jégou, Douze M , Schmid C . Packing bag-of-features. EEE International Conference on Computer Vision. IEEE 2010.
[51] Li J , Qian X , Tang Y Y , et al. GPS estimation for places of interest from social users' uploaded photos. IEEE Transactions on Multimedia 2013, 15(8):2058-2071.
[52] Li, Xinchao, Larson, et al. Global-scale location prediction for social images using geo-visual ranking. IEEE Transactions on Multimedia 2015.
[53] Deretey E , Ahmed M T , Marshall J A , et al. Visual indoor positioning with a single camera using PnP. 2015 International Conference on Indoor Positioning and Indoor Navigation (IPIN). IEEE 2016.
[54] Qian X , Hua X S , Chen P , et al. PLBP: An effective local binary patterns texture descriptor with pyramid representation. Pattern Recognition 2011,44(10-11):2502-2515.
[55] Guo Z , Zhang L , Zhang D . Rotation invariant texture classification using LBP variance (LBPV) with global matching. Pattern Recognition 2010, 43(3):706-719.
[56] Hussain, S.; Triggs, B. Visual recognition using local quantized patterns. In Computer Vision ECCV 2012; pp. 716–729.
[57] Qin T, Li P, Shen S. Vins-mono: A robust and versatile monocular visual-inertial state estimator. IEEE Transactions on Robotics 2018, 34(4): 1004-1020.
[58] Mur-Artal R, Montiel J M M, Tardos J D. ORB-SLAM: a versatile and accurate monocular SLAM system. IEEE transactions on robotics 2015, 31(5): 1147-1163..
[59] Leutenegger S , Furgale P , Rabaud V , et al. Keyframe-based visual-inertial SLAM using nonlinear optimization. Proceedings of Robotics: Science and Systems 2013.
[60] Balntas V, Li S, Prisacariu V. Relocnet: Continuous metric learning relocalisation using neural nets. Proceedings of the European Conference on Computer Vision (ECCV) 2018: 751-767.
[61] Chiou M J, Liu Z, Yin Y, et al. Zero-shot multi-view indoor localization via graph location networks. Proceedings of the 28th ACM International Conference on Multimedia. 2020: 3431-3440.