Noname manuscript No.
(will be inserted by the editor)

A Review of Monocular Visual Odometry

Ming He · Chaozheng Zhu · Qian Huang · Baosen Ren · Jintao Liu · QUINTAN UVIA

Received: date / Accepted: date

Abstract Monocular Visual Odometry provides more robust functions on navigation and obstacle avoidance for mobile robots than other Visual Odometries, such as binocular Visual Odometry, RGB-D visual and basic odometry. This paper describes the relationships between Visual Odometry and Visual SLAM (simultaneous location and mapping). The basic principle of Visual Odometry is expressed in the form of mathematics, specifically by incrementally solving the pose changes of two series frames, and further improving the odometry through global optimization. After analyzing the three main ways of implementing visual odometry, the state-of-the-art monocular Visual Odometries, including ORB-SLAM2, DSO and SVO are also analyzed and compared in detail. The issues of robustness and real-time opetation,which are generally of interest in the current visual odometry research are discussed from the development directions and trends. Furthermore, we present a novel framework for the implementation of next-generation Visual Odometry based on additional high dimensional features, which has not been implemented in the relevant applications.

Keywords Monocular Visual Odometry · Feature Method · Direct Method · visual SLAM

Ming He
College of Command Control Engineer, Army Engineering University at Nanjing
E-mail: 1091721005@qq.com

Chaozheng Zhu
College of Command Control Engineer, Army Engineering University at Nanjing
E-mail: 370045744@qq.com

Qian Huang
College of Information and Computer, HoHai University at Nanjing
E-mail: huangqian@hhu.edu.cn

Baosen Ren
State Grid ShanDong Electric Power Maintenance Company at Linyi
E-mail: 18354261031@163.com

Jintao Liu
College of Command Control Engineer, Army Engineering University at Nanjing
E-mail: top2012@163.com

QUINTAN UVIA
the Papua New Guinea Defence Force
E-mail: 1706020113@hhu.edu.cn

1 Introduction

Positioning and navigation of mobile robots in unknown environments is an essential function in autonomy. Due to the complexity of the unknown environment, it is of great significance to build a real-time map and locate the identify poses based only on the robot's own sensor [1] [2]. Visual sensors,a common type of robot sensor,have the advantages of high accuracy, low cost and abundant data information. Therefore, using a visual sensor to determine location has become a main topic of research. The concept of visual odometry [3], proposed by Nister, refers to using machine vision technology and a related image sequence analysis to estimate the mobile robot pose (position and attitude) in real time. This process can also overcome the shortcomings of traditional odometry and provide more accurate positioning. Furthermore, it can run where the Global Position System (GPS) is not available, such as indoor environments or interplanetary exploration [3] [4].

In view of the features and advantages of the Visual Odometry (VO), it has been successfully applied to the Mars exploration [4]. It also highlights its important application value in the fields of public security, Virtual

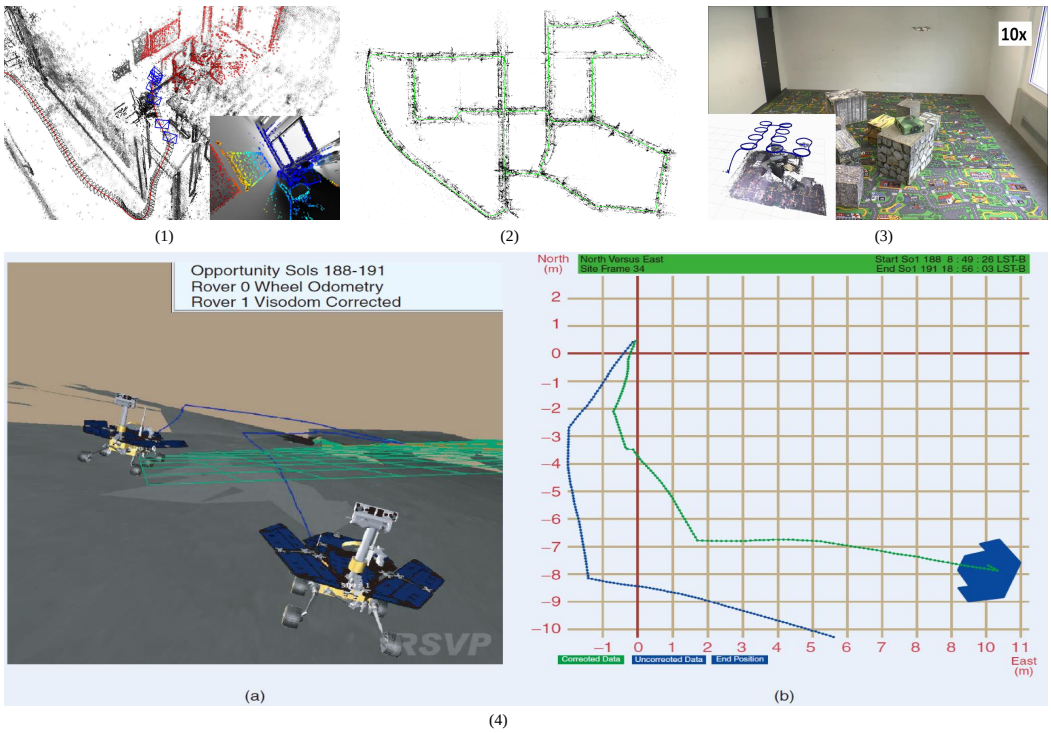

Fig. 1 Research and application of VO samples, where (1), (2), and (3) represent the state-of-art research on VO in the form of DSO, ORB-SLAM2 and SVO, respectively. (4) is a real application in 2004: the use of visual odometry on Mars.

Reality (VR) [5], Augmented Reality (AR) [6] and so on. As shown in Figure 1.

1.1 Difference and connection between visual SLAM and VO

There are two mainstream methods based on visual Simultaneous Location and Mapping (vSLAM) [7]. One is a general approach to apply classical filter [8] to vision information fusion, the other to exploit selected key frames to develop global optimization [9,10]. The detailed evaluation of these two approaches is described in [11,12].

The difference between vSLAM and VO is that the latter focuses only on the consistency of local trajectories, while the former focuses on the consistency of the global trajectory. Understanding when to generate a loop and effectively integrate new constraints to the current map is a major research issue in visual SLAM. The VOs target is an incremental reconstruction trajectory, which may only optimize the position of the first n paths, which is termed sliding-window-based bundle adjustment. The sliding window-based optimization relies on a local map in SLAM.

The connection between vSLAM and VO is that the latter can be regarded as a module within the former and can reconstruct the trajectory of the camera incrementally. Thus, some scholars regard vSLAM as VOs expansion research. In terms of application scenarios, VO is sufficient in such situations where real-time localization is needed, such as missile guidance flights and UAV and AR scenarios. However, this aspect is redundant for vSLAM to build an accurate map, which could waste additional computing power.

In recent years, significant progress [13–16] has been made in both monocular and binocular cameras. Most of these devices can operate in wide range of outdoor environments. As shown in Table 1 [17], since Parallel Tracking and Mapping (PTAM) was implemented in 2007, due to the special structure of the sparse matrix, the back-end research has progressed from EKF to optimization [18,19].

2 Formal Description of VO

For a monocular VO, at time k, the image sets designated $I_{0:n} = I_0, ..., I_n$, are collected by the camera of a rigid robot. Suppose that the camera coordinates are the coordinates of the robot. In the stereo vision system, the left camera is typically the original. As shown in Figure 2.

However, the use of binocular VO leads to a sharp decline in the accuracy of triangulation as the distance

Table 1 Classic VO research results

Solution name	Publish time	Sensor type	Implementation method	Back-end Optimization method	Characteristic
MonoSLAM [18]	2007	Monocular	Feature point method	EKF Filter	First visual SLAM in real-time, EKF + sparse features
PTAM [19]	2007	Monocular	Feature point method	Optimization	Keyframes + BA; first-time use of optimization as the back end
DTAM [25]	2011	Monocular	Direct method	Optimization	Direct method, monocular dense map, needs GPU support
Kinect Fusion [30]	2011	RGB-D	Direct method	Optimization	First implementation of dense reconstruction based on RGB-D in real-time
RTAB-MAP [24]	2013	Binocular/RGB-D	Feature point method	Optimization	Supports larger scene
DVO [26]	2013	RGB-D	Direct method	Optimization	Direct method based on RGB-D, dense map
SVO [20]	2014	Monocular	Semi-direct method	Optimization	Sparse direct method, only VO
LSD-SLAM [22]	2014	Monocular	Direct method	Optimization	Direct method + semi-dense map
RGB-D-SLAM-V2 [7]	2014	RGB-D	Feature point	Optimization	Complete RGB-D dense reconstruction
OKVIS [9]	2015	Monocular/multi-cameras + IMU	Feature point method	Optimization	Mainly optimization based on key-frame VIO
ROVIO [8][29]	2015	Monocular +IMU	Direct method	EKF Filter	Mainly EKF based on VIO
Elastic Fusion [27][28]	2015	RGB-D	Direct method	Optimization	RGB-D reconstruction in real-time, visualization
DSO [21]	2016	Monocular	Direct method	Optimization	Monocular direct method,best results of the direct method at present
ORB-SLAM2 [16][23]	2017	Mainly monocular	Feature point method	Optimization	ORB feature + three thread structure
VINS-Mono [10]	2017	Monocular + IMU	Feature point method	Optimization	Tightly-coupled framework of VIO based on optimization

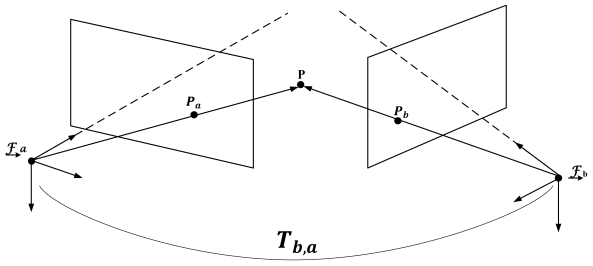

Fig. 2 The illustration of the VO problem.

between the center of the two cameras is affected by the conditions of the measurement accuracy and climate change (such as, thermal expansion and cool contraction). Therefore, this paper focuses on the research of the monocular VO problem.

Two camera poses near and times k and $k-1$ form a rigid transformation $T_{k,k-1} \in \mathbb{R}^{4\times4}$ as follows:

$$T_{k,k_1} = \begin{bmatrix} R_{k,k-1} & t_{k,k-1} \\ 0 & 1 \end{bmatrix} \tag{1}$$

where $R_{k,k-1}$ is a rotation matrix, and $t_{k,k-1}$ is a translation matrix. Set $T_{1:n} = \{T_{1,0}, T_{2,1}, ..., T_{n,n-1}\}$ contains all the motion sequences. Finally, the camera position set $C_{0:n} = \{C_0, C_1, ..., C_n\}$, where C_n is the initial coordinates at time k. The current position C_n can be calculated by the connection between the transformation $T_k(k = 1n)$, therefore, $C_n = C_{n-1}T_n$, C_0 is the camera pose at time $k = 0$.

The main goal of VO is to calculate T_k from the image I_k to the image I_{k-1} and then integrate all the transformations to restore the entire path $C_{0:n}$ of the camera. Here, VO is a position followed by a position or an incremental reconstruction trajectory. An iterative optimization based on the front m poses can be executed, and then a more accurate local trajectory estimation can be obtained.

Iterative optimization minimizes the reprojection error of 3D points in the local map based on the former m frames (based on window binding adjustment because it executes on m frame window). The depth of 3D points in the local map space is estimated by triangulation. Therefore, an optimization problem can be constructed, adjusting R, t, so that for all feature points z^j, the cumulative error of the two norms is minimal, and the results are as follows:

$$\begin{aligned} min_{X,R,t} \sum_{j=1}^{N} \left\| \frac{1}{\lambda_1} CX^j - \left[z_1^j, 1 \right]^T \right\|^2 + \\ \left\| \frac{1}{\lambda_2} C(RX^j + t) - \left[z_2^j, 1 \right]^T \right\|^2 \end{aligned} \tag{2}$$

This is the problem of minimization of the reprojection error. In the actual operation, each X^j is adjusted to increase consistency with each observation z^j and to minimize every error term as much as possible. For this reason, it is also called Bundle Ad-justment. The principle of bundle adjustment and optimization is shown in Figure 3.

3 Research Progress of The VO Method

Over the past 10 years,VO [22,31] of large-scale scenes has achieved great success. The method of VO implementation includes the feature point method and direct method. The hybrid semi-direct tracking method is a combination of both methods.

3.1 Method based on feature points

For the feature point method [3,6,16,18,23,32,33], Nister was the first to carry out real-time monocular large scene VO related work [3].VO of sparse feature points is the current mainstream method [32,34]. The basic idea is for every frame of new image I_k (a pair of images in a stereoscopic camera), the first two steps are to detect and match 2D feature points and match them with the previous frames. The reprojection of two-dimensional feature points is the extraction of common 3D feature points from different image frames, which provides the corresponding relationship of images (most VO implementation assumes that the camera has been calibrated).

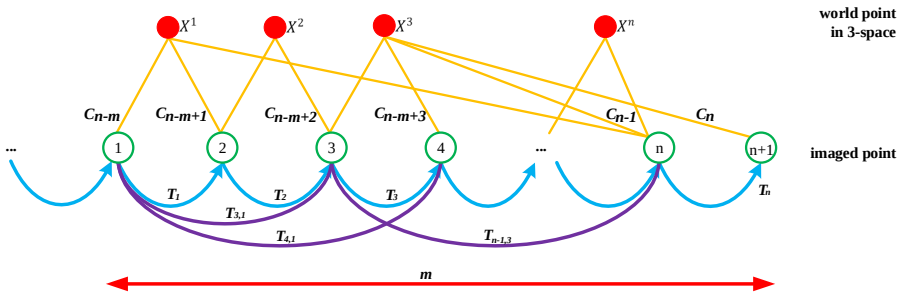

Fig. 3 Principle of bundle adjustment and optimization, where C represents the camera pose of the current frame, T represents the transformation of the pose between the two cameras and m represents the total number of cameras.

The third step is to calculate the relative motion T_k between k-1 and k. According to the corresponding relationship is three-dimensional or two-dimensional, there are three different methods, including 2D-2D, which can be used to solve this problem, including the polar constraint, P3P (Perspective-Three-Point) and ICP Iterative Closest Point) [35–38]. The position and posture of the camera C_k is based on the transformation of the previous position T_k. To achieve more precise local trajectory estimation by iterative optimization (bundle adjustment), we build a local map based on the depth estimation of the previous m frames. Figure 4 shows a flow chart of the VO system based on the feature point method.

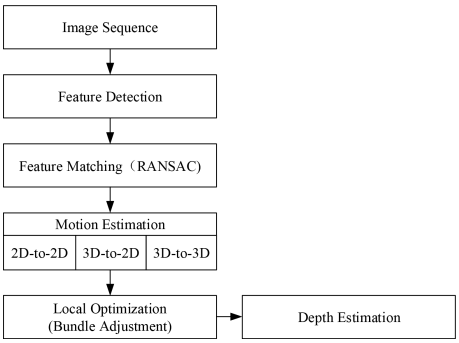

Fig. 4 Main flow chart of the VO system based on the feature point method.

In addition, we focus on the noise, erroneous measurements and erroneous assumptions on the data, which tend to lead to matching outliers in the process of feature matching. Even in the case of outliers, robust estimation is required to ensure accurate motion estimation. Because of the decentralized nature of the outliers, the Random Sampling Consistency (RANSAC) is used to select the optimal matching, but not the least square matching algorithm.

The main research problem of VO is how to estimate the motion of the camera according to the image. Typically, due to the influence of illumination and deformation on the gray value, the change between different images may be very large. Therefore, only the gray value is insufficient; we need to extract feature points from the images. In the context of computer vision and image processing, a feature is a group of related information and computing tasks depending on the application. The feature may also be the result of feature detection or a general neighbourhood operation applied to the image. Features may have special structures in the image, such as corner points, edges, or block objects [39]. However, it is generally easier to find the same corner in the two images, whereas finding the same edge is slightly harder, and finding the same block is the most challenging. Therefore, an intuitive method of feature extraction is to identify the corner points of different images and determine their corresponding relationship. In this case, the corner point is the so-called feature.

However, in the actual situation, the simple corner point still cannot meet the requirements. Consequently, researchers have designed multiple additional stable local image features, such as SURF (Speeded Up Robust Features) [40] and SIFT (Scale-Invariant Feature Transform) [41]. Although SIFT and SURF fully consider all kinds of problems in the process of image transformation, they also incur a large amount of computation. Generally, it is challenging to execute calculation in real time on a CPU. However, in recent years, the popularity of some computable feature extraction/description algorithms, such as ORB (Oriented FAST and Rotated BRIEF) [42] and BRISK (Binary Robust Invariant Scalable Keypoints) [43], has gradually exceeded that of the Harris corner points or SIFT/SURF, which were not well tracked before, and the former group of algorithms are now preferred in VO.

The ORB combines the advantages of FAST (Features from Accelerated Segment Test) [4] and BRIEF (Binary Robust Independent Elementary Features), providing strong features in scale, rotation, and brightness, for example. Moreover, the combination is very efficient, making the ORB the best current real-time scheme [16]. Typically, features are made up of key points and descriptors. Among them, consider FAST corner extraction: ORB increases rotation invariance in the descriptor and increases the main direction of the feature points on the basis of the original FAST. Consider also the new BRIEF descriptor: to describe the pixel area around the key points extracted before, because the main direction is added when the corner points are extracted, the descriptors of ORB have better rotation invariance than that of the original BRIEF [45] descriptors.

This paper mainly compares three main methods of feature point extraction, namely, SIFT, SURF and ORB, all of which have been implemented in OpenCV, as shown in Table 2.

Table 2 Performance comparison between different features

	Feature type		
	ORB	SURF	SIFT
Complexity	√√√√√	√√√	√
Rotational robustness	√	√√√	√
Blur robustness	√	√√√	√
Scale transformation robustness	×	√√	√

Early real-time VO was based on the feature point; see, for example, the monocular VO framework (PTAM [19]) proposed by Klein et al. Although its performance is not efficient, this approach provides a complete and universal framework for the implementation of visual odometry. With respect to the realization of Visual Odometry, this process can be divided into front-end and back-end, parallel processing tracking and mapping tasks. Most of the VO frameworks are based on this implementation, including the most stable second-generation simultaneous location and mapping based on ORB (ORB-SLAM2) [16]. It is also the first system to use nonlinear optimization. Traditionally, the implementation of VO is based on the filter [18].However, there are some disadvantages involving the small scene and lack of global relocation, resulting in poor applicability.

The optical flow method has the characteristics of feature point tracking. This method is superior to other feature point matching methods in that it can reduce calculation somewhat, so there is a visual odometry system called Flowdometry,it is proposed based on optical flow and deep learning. Optical flow images are used as input to a convolutional neural network, which calculates a rotation and displacement for each image pixel. The displacements and rotations are applied incrementally to construct a map of where the camera has traveled.

The most useful feature-based VO method in existing research is ORB-SLAM2 [16], which presents a more complete VO framework, as shown in Figure 5. This method includes tracking, mapping and loop detection of three threads. Among these methods, the tracking thread is mainly responsible for extracting the ORB [42] feature points for a new frame image and roughly estimating the position and posture of the camera. The mapping thread is mainly based on bundle adjustment to optimize the feature points and camera pose in the local space so that the space position of the feature points with smaller errors is solved. The loop detection thread is responsible for the realization of loop detection based on the key frame, which can effectively eliminate the accumulative error and can also carry out global reposition. This scheme is also compatible with monocular, binocular and RGB-D cameras, making it well suited for general use.

For initialization, [16] proposes an strategy for automatic initialization map, and calculates the homography matrix (assuming a planar scene) [31] and essential matrix (assuming non-planar scene) [32]. According to the heuristic rule, the corresponding situation is determined to initialize the pose and position. This is also the most greatest contribution in document [16]. The computing advantages of ORB-SLAM and PTAM are not only the more efficient ORB features selected but also the matching points that can be observed on the previous frame rather than directly using all map points to match the new frames.

3.2 Method based on direct tracking

The feature point method has long been a widely used method, but its robustness is mainly based on the description of the feature points. On the one hand, the robustness is enhanced, and the complexity of the feature point description is increased, leading to a large increase in the complexity of the algorithm. However, feature points cannot be applied to scenes with weaker feature points, as is the case with metopes and in Skype. Therefore, the direct method of estimating camera motion based on the pixel gray invariance hypothesis has developed rapidly in recent years [20] [21]. The direct method,which is developed from the optical flow [48], can estimate the camera motion and the pixels spatial location by minimizing the photometric error (minimizing the reprojection error of feature points in the fea-

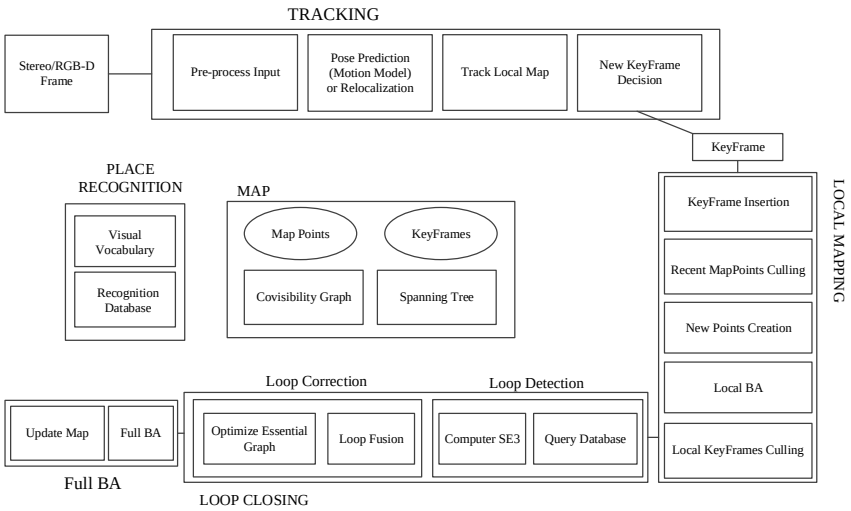

Fig. 5 ORB-SLAM2 frame structure chart

ture point method) without addressing the feature (or not calculating the feature description). This approach can effectively solve the problems faced by the feature point method. In general, the direct method is divided into three categories according to the space point P, the sparse direct method, the semi-dense direct method and the dense direct method.

The early direct VO method was rarely based on the tracking and mapping framework, most of which involved the key points of artificial selection [49–51]. Upon the appearance of the RGB-D camera, researchers found that the direct method was very effective for this camera type [26] and then for the monocular camera [21,22]. Recently, direct methods have appeared that directly use the image pixel gray information and geometric information to construct the error function through the graph optimization to minimize the cost function, thus obtaining the optimal camera pose. These methods are applied to large-scale map problems with map pose and position graph [21,52]. To construct a semi-dense 3D environment map, Engel et al. [22] proposed the large-scale direct monocular simultaneous localization and mapping (LSD-SLAM) algorithm to replace the previous direct methods of VO. This method enables high-precision estimation of the camera pose and position to create a large-scale 3D environment map. Because monocular VO suffers from scale uncertainty and the scale drift problem, the map is directly composed of a key frame direct Sim(3) transformation, which can detect scale drift accurately, and the whole system can run on a CPU in real time. Similar to ORB-SLAM2, LSD-SLAM is also optimized with pose and position graph, so it can form a closed loop and accommodate large-scale scenarios. The system selects the nearest key frame position for each newly added key frame in the existing key frame set (map). The main flow chart of LSD-SLAM is shown in Figure 6.

DSO[21] (direct sparse odometry) was also proposed by Engel, the inventor of LSD-SLAM. DSO improves the robustness, accuracy, and speed of computation, surpassing previous ORB-SLAM and LSD-SLAM methods. Here, the new depth estimation mechanism is used to optimize the sliding window instead of the original Kalman filtering method, providing a substantial improvement in accuracy. In addition, in contrast to LSD-SLAM, DTAM [25] provides a direct method to calculate a real-time dense map based on a monocular camera. The pose and position estimation of the camera uses a depth map to directly match the whole image. However, computing dense depths from a monocular vision requires substantial computing power, typically using GPU parallel operations, such as open source REMODE [53].Multiple researchers have worked in this area to ensure faster computing speed, as in [54] and [20].

3.3 Method based on the hybrid semidirect tracking

Although the method based on direct tracking is very widely used, a low speed and lack of assurance of optimality or consistency are problems of the direct method. Therefore, based on the advantages of the feature-based method and direct tracking method, a hybrid semidirect method is proposed, namely, semidirect visual odometry [20]. Although SVO is still dependent on the characteristics of consistency, this method applies the direct

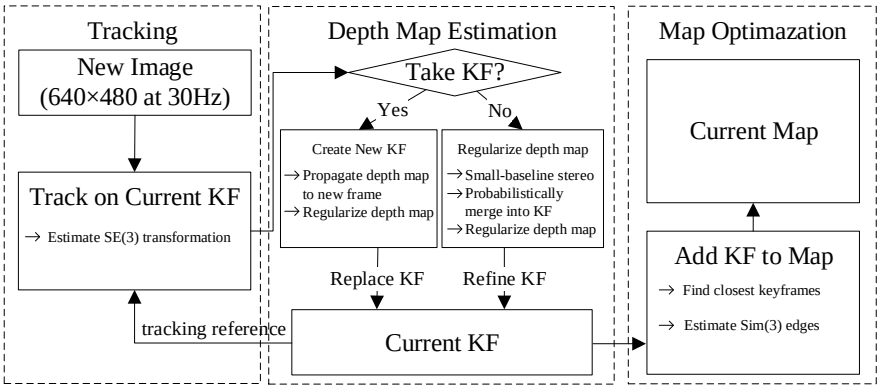

Fig. 6 Module flow chart of LSD-SLAM

method to obtain the pose and position. This approach can help eliminate the feature matching and peripheral point processing to greatly shorten the calculation time. The algorithm is very fast. 55 fps can be achieved on the Embedded UAV platform (ARM Cortex A9 1.6 GHz CPU), and the framerate can be as high as 300 fps on a general laptop (Intel i7 2.8 GHz CPU).

Depth estimation is the core of building a local point cloud map, and SVO is also built with a probability model. However, unlike LSD-SLAM or other methods, the deep filtering of SVO is based on a mixed model of Gauss distribution and homogeneous distribution [55], while LSD-SLAM is based on the Gauss distribution model. First, the direct method is used to solve the pose and position matching. Second, classical Lucas-Kanade optical flow [48] matching is used to obtain subpixel accuracy. Then, the minimized reprojection error is optimized by combining the point cloud map, as shown in Figure 7.

In contrast to using the traditional feature points, the whole process needs to rely on features when selecting key frames only. The calculation of matching descriptors is removed, and the steps of using the RANSAC to remove the outliers are removed, so the process is relatively efficient; relative to the direct method, this method does not directly match the whole image to obtain the pose and position of the camera. Instead, it extracts the image block from the whole image, allowing us to obtain the pose and position from the image block. This technique enhances the robustness of the algorithm. The largest contribution of SVO is the ingenious design of the three optimization methods (i.e., optimize the gray error, optimize the feature point prediction position, and optimize the reprojection error) to meet the accuracy problem while maintaining excellent computing speed. In addition, its code structure is relatively simple and very suitable for further study. Forster proved that this method could be extended to the multiocular system [56], tracking the edge, including the prior knowledge of motion. The method also supports various cameras, such as fish-eye and perspective cameras. In the field of semidirect research, however, there are no influential papers except for Forsters latest paper [56].

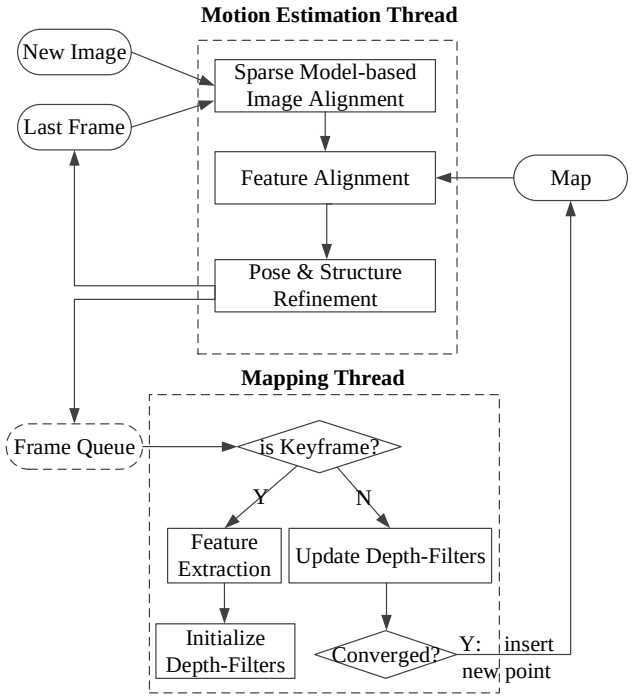

Fig. 7 Module flow chart of SVO

3.4 Analysis of advantages and disadvantages

The feature point method has always been a classical method, but its robustness is mainly based on the description of feature points. On the one hand, robustness is enhanced and the complexity of the feature point description is increased while the complexity of the algorithm is greatly reduced. On the other hand, feature points cannot be applied to scenes with weak feature points, such as walls. The direct method is a relatively new method that can be adapted to scenes with insufficient features, such as corridors or smooth walls [57],and has strong robustness. By skipping the feature description and matching steps, the direct method, particularly the sparse direct method, tends to run at extremely high speeds. The method is also compatible with requirement scenarios that need to build a semi-dense map or dense map, which is not possible using the feature point method. However, there are also some problems such as nonconvexity, single-pixel nonsegmentation and the poorly supported assumption of gray invariance in the direct method; thus, its research and implementation are not as mature as that of the feature point method. At present, the direct method is suitable only for small motion situations and small overall brightness changes. Although VO [16,23] based on feature points is more mainstream, from the experimental results published by TUM group of the University of Munich, the direct method of VO [20–22] has made great breakthroughs in recent years. Among them, the sparse direct method [21] has a faster and better than the sparse feature point method [16]. The direct method uses all the information on the image, or even a small area of the pixel gradient, so even in the case of poor scene texture, the performance of the focus and motion blur is better than that of the feature-based method. According to a comparison of noise experiments based on the direct tracking method and the feature-based method [21], the direct tracking method is more sensitive to geometric noise, such as that produced by a rolling shutter camera. Features-based methods are more sensitive to optical noise, such as fuzzy noise. Therefore, on ordinary mobile devices (typically the shutter camera), the feature-based method may work better. In the robot based on the global shutter camera, the method based on direct tracking is becoming increasingly widely used. A method based on hybrid semi-direct tracking [20], first proposed by Forster, has the advantages of fast speed and suitability for the map uncertainty model and is not affected by the assumption of model motion. However, due to fewer tracking features, some cases may be lost. Besides, the author released an amazing experimental video, and open the source code of implementation framework. Although its open source code is not very robust, this method is still well suited for beginners to study because of its straightforward code implementation. To better understand the progress of the comparison of the various methods, this paper evaluates the most representative method of the feature-based method, the method based on direct tracking, and the method based on hybrid semi-direct tracking through experiments. The results are shown in Figure 8.

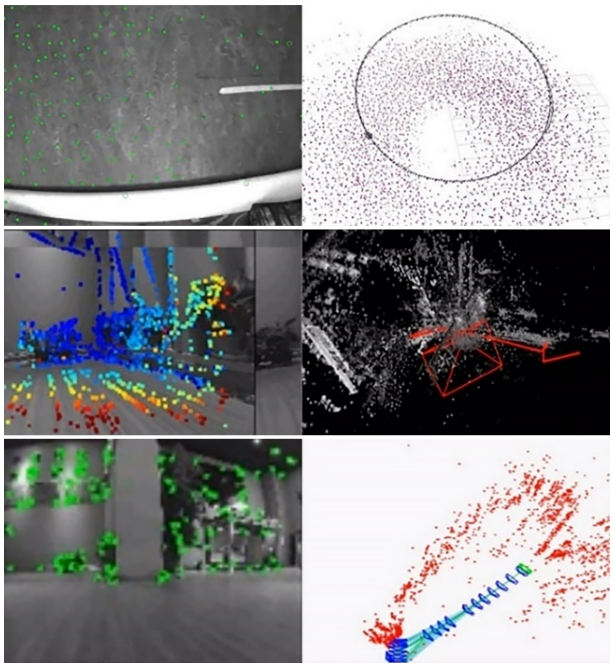

Fig. 8 Comparison of the effect of the three types of VO methods. Top line: SVO algorithm; middle line: DSO algorithm; bottom line: ORB-SLAM2 algorithm.

4 Main development trends and active research areas of VO

Table 3 shows the academic research institutions worldwide that have contributed greatly to VO.

Although the VO problem is essentially an incremental calculation of the pose and position of the camera, this technique provides a pose and position estimation for the upper-level applications itself. How to further improve accuracy, efficiency and robustness remains a persistent aim of researchers. Around the above three problems, there have been several active areas of research, such as exploring new sensors, multisensor data fusion, applied machine learning, exploring new mitigation feature dependency and reducing computational complexity.

Table 3 research direction of frontier institutions all over the world

Research Affiliation	Research Directions
University of Zurich	Direct method, VO based on a novel visual sensor
University of Munich	Direct method
National Aeronautics and Space Administration	Binocular stereo visual odometry
The Hong Kong University of Science and Technology	Visual-inertial fusion
Apple	Visual-inertial fusion
Google	RGB-D camera and inertial fusion
Swiss Federal Institute of Technology Zurich	Visual-inertial fusion
Tsinghua University	Semantic map reconstruction in machine learning
Zhejiang University	Binocular stereo visual odometry
SZ DJI Technology	Binocular stereo vision system with inertial measurement element
MI	Laser vision-multiple sensors fusion

4.1 Explore new sensors

Microsofts RGB-D camera Kinect (released in 2010) can obtain a depth map in real time and simplify calculations substantially, enabling the realization of a dense 3D reconstruction system [7,24,26–28,30]. However, due to its short effective distance, susceptibility to interference by external light sources and incompatibility with outdoor scenes, Kinect is not the ultimate solution to the VO problem. In recent years, event-based cameras have attracted research attention. The advantages of event-based cameras with respect to standard cameras are their low latency, high dynamic range, low bandwidth, and low power, for example. Such novel cameras require new algorithms to address the problems of no-intensity information and very low image resolution, however. In 2017, Zihao Z et al. of the University of Zurich proposed a VO algorithm based on the event camera. Moreover, based on the extended Kalman filter and the unstructured measurement model, IMU was integrated as a complement to data fusion to accurately obtain the pose and position of University of Zurich proposed VO algorithm based on event camera. Moreover, based on the extended Calman filter and the unstructured measurement model, IMU is integrated as a complement to data fusion to get the pose and position [58] of a 6 DOF camera. The emergence of new sensors is expect to generate comparable research interest.

4.2 Multisensor data fusion

For many mobile robots, IMU and vision are necessary sensors, as they can complement each other by data fusion to meet the need for mobile robot system robustness and location accuracy. The combination of monocular camera and inertial navigation [8–10,31,59] has also been a notable trend in recent years. Apple Inc.'s ARKit, launched at the WWDC 2017 conference, is mainly based on the idea of EKF for a monocular camera and inertial navigation data fusion, providing a solid foundation platform support for developers to implement indoor positioning. Later, it was proposed to [9] integrate multiocular and inertial navigation data with the optimized key frame [60].Data fusion is divided into tight coupling and loose coupling. On the one hand, to limit the computational complexity, much work has followed the principle of loose coupling. One study [31] integrated IMU as an independent attitude and related yaw measurement to address the nonlinear optimization problem of vision. In contrast, another study [61] used visual pose and position estimation to maintain an EKF of an indirect IMU. Similar loose coupling algorithms include [62] and [63];here, the pose and position estimation of the camera uses a nonlinear optimization set to the factor graph, including inertial navigation and GPS data. On the other hand, the loose coupling method essentially neglects the correlation between different sensors. The tightly coupled method combines camera and IMU data and jointly estimates all states as a common problem, so we need to consider the correlation between them. A previous report [9] compared these two methods. Experiments show that the correlation between these sensors is very critical for the high-precision visual inertial navigation system (VINS), so the high-precision visual inertial navigation system is tightly coupled. Many researchers have explored multi-sensor fusion, e.g., the integration of multicamera sensors [64] proposed by Yang Shaowu with binocular stereo vision and inertial navigation, speed and data fusion [65]. Second, Akshay proposed a GPS-Lidar fusion algorithm based on a point cloud feature, which can effectively reduce the position measurement error in 3D urban modeling [66].

4.3 The application of machine leaning

In recent years, machine learning methods such as neural networks have caused a widespread academic sensation in many fields, and the VO field is no exception. In the matching tracking part, a data-driven model

(3DMatch) was proposed [67].The local spatial block descriptor is obtained from the existing RGB-D reconstruction results by self-supervised feature learning; then, the corresponding relationship between local 3D data is established. For optimization of matching errors, traditional RANSAC can be replaced by a new highway network architecture. This approach is based on multilevel weighted residual shortcuts and every possible parallax value calculation of matching error and by using composite loss function training as a support for multiple comparisons of an image block. This framework can be used to better detect the exception points in the refinement step. Previous experiments [68] on this new architecture using the stereo matching datum data set showed that the matching reliability is far superior to the existing algorithms.

The lack of scale information in monocular VO has always been the issue of greatest concern for researchers. Recently, German researchers, such as Keisuke et al., addressed scaling failures for monocular VO areas such as low-texture areas. Keisuke et al. proposed a fusion method involving depth information predicted by a CNN and depth information directly calculated by a monocular process. The technique was experimentally shown to solve the scale information loss problem of monocular VO [69].

In 2018, Ruben Gomez-Ojeda et al. proposed a method called Learning-based Image Enhancement for Visual Odometry in Challenging HDR Environments. It also propose a convolutional neural network of reduced size capable of performing faster and overcome one of the main open challenges in VO is the robustness to difficult illumination conditions or high dynamic range (HDR) environments. And a novel monocular VO system called UnDeepVO and a novel end-to-end framework for monocular VO by using deep Recurrent Convolutional Neural Networks (RCNNs) perform better other monocular VO methods in terms of pose accuracy.Therefore, the application of machine leaning will become the next hot spot in the VO field.

4.4 Mining high-dimensional information in Vision

The dependency of VO on scene features is essentially due to the use of the overly underlying local features (point features). Therefore, multiple methods have been proposed to reduce the feature dependence by using image information, such as edge and plane information [71]. In theory, the edge can carry information such as the direction, length, and gray value, so it is more robust. The edge-based features in the indoor scene (more regular items) are expected to provide better robustness. One study [72] proposed a monocular VO algorithm that combines point and edge advantages. This algorithm not only performs well in the monocular open dataset [21] provided by TUM but also greatly reduces the motion estimation error in low-texture environments. Study [Direct Line Guidance Odometry] proposed an extension to a point-based direct monocular visual odometry method, It uses lines to guide keypoint selection rather than acting as features. Thereby increasing efficiency and accuracy. Another study [73] mainly applied graph model and graph matching mechanism to track planar objects and designed a new strategy to solve optimal problems, which can predict the posture and key point matching of objects.

4.5 Reduce computational complexity

At present, the real-time recovery of dense scenes based on RGB-D cameras has been improved [7, 24, 26–28, 30]. In recent years, due to the outbreak of the AR application research, Schops et al. at Google have developed an early foundation for AR technology support, proposing to apply TSDF to fusion depth maps to realize 3D reconstruction methods [74] in the so-called Project Tango. The main computational complexity results from an excessive number of depth estimates needed for semi-dense or dense reconstruction. Consequently, most of the optimization methods at present are focused on optimizing depth estimation steps. For example, DTAM [25] introduces regularization terms to global optimization of the depth map to reduce the probability of error matching. In REMODE [53], the depth filter model is used to optimize the parameters of the update probability model of each frame. Although these methods can reconstruct the dense 3D point cloud map in real time, most of them still rely on the parallel acceleration operation of the GPU. In 2017, Euntae Hong proposed a novel visual inertial odometry algorithm which directly optimizes the camera poses with noisy IMU data and visual feature locations. The proposed algorithm is conceptually very clear and simple, achieves good accuracy. However, how to improve the computational efficiency and reduce the computational complexity remains a topic of interest to recover the monocular semi-dense or dense 3D point cloud map based only on a CPU.

4.6 A new novel framework of VO

The ORB feature contains 4DoF information including scale invariance (z), rotation invariance (θ), and translation invariance (x, y). Some scholars have studied the angle optical flow based on Oriented FAST to localize

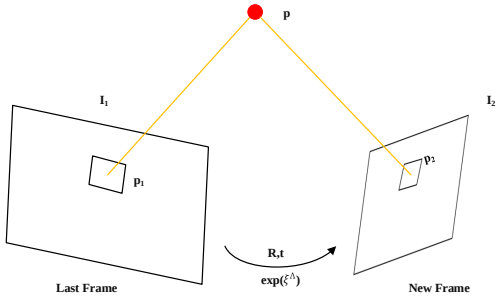

Fig. 9 Direct high-dimensional optical flow diagram

the 3DoF of a camera (x, y, θ). In contrast to previous classical LK optical flow,in which only 2DoF of a camera can be obtained [48], it is possible to provide a 1DoF increase to a camera with the improved corner feature. Similarly, it will soon be possible to directly describe the corner with the features of a simple, higher dimension. This method will be combined with the theory of intensity invariance and nonlinear least squares to try to solve the 6DoF problem for a camera, including rotation and translation, and then incrementally solve the VO problem, as shown in Figure 9. By eliminating the need for costly feature matching and decomposing the essential matrix (as required by conventional methods), this capability is expected to sharply reduce the algorithmic complexity. On the other hand, the direct method requires less motion for two frames. When the complexity of the algorithm is reduced, the frame rate is increased significantly (e.g., in some special cases such as a mouse, the optical flow based on the theory of intensity invariance is used to solve for its position, and the frame rate can exceed 1800 fps). This scheme can further improve the accuracy of VO.

5 Conclusion

This paper starts with the comparative analysis of VO and visual SLAM and formalizes the VO problem. Then, we focus on the research progress of various methods to realize VO and compare and analyze the advantages and disadvantages of these methods. Finally, along with discussing the research directions of prestigious scientific research institutions worldwide, we summarize the development of future active research areas. At present, most researchers focus only on ideal scenes with a satisfactory visual field, such as daytime, but a variety of both indoor and outdoor scenes (from day to night and with seasons changing, for example) are very common. How to ensure that the VO system remains highly robust under such circumstances is an important research direction. In addition, to satisfy real-time performance demands, we present the concept of novel angle optical flow, which may decline the complexity of VO.

In addition to traditional application fields such as drones and unmanned vehicles, in the future, new application research might be conducted in the following areas. In the field of firefighting, in the harsh environment of large-scale indoor fires, it is essential for firefighters to be positioned and rely on motion trajectory in real time. Search-and-rescue work efficiency benefits from marking the places that have been searched. Finally, in the field of counter-terrorism, VO-based police dog pose estimation can provide counter-terrorism personnel with a low-cost, over-the-horizon, interactive approach.

Acknowledgment

This work was supported by National Key R&D Program of China Nos. 2018YFC0806900, 2016YFC0800606, 2016YFC0800310, and 2018YFC0407905; Natural Science Foundation of Jiangsu Province under Grants Nos. BK20150721 and BK20161469; Primary Research & Development Plan of Jiangsu Province under Grants Nos. BE2015728, BE2016904, BE2017616, and BE2018754.

References

[1] H. Durrantwhyte,T. Bailey.*Simultaneous localization and mapping:Part I.* IEEE Robotics & Automation Magazine,13(3): 108-117 (2006).

[2] H. Durrantwhyte,T. Bailey.*Simultaneous localization and mapping:Part II.* IEEE Robotics & Automation Magazine,13(3): 108-117 (2006).

[3] D. Nister,O. Naroditsky,J. Bergen.*Visual odometry* In Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition,652-659 (2004).

[4] L. Matthies,.M Maimone,A. Johnson,et al.*Computer vision on mars.*International Journal of Computer Vision,75(1): 67-92 (2007).

[5] C. Malleson,.A Gilbert,M. Trumble,et al.*Real-time full-body motion capture from video and IMUs* In Proceedings of International Conference on 3DVision (2017).

[6] T. Wang,H. Ling.*Gracker:A graph-based planar object tracker.* IEEE Transactions on Pattern Analysis & Machine Intelligence, 99 (2017).

[7] F. Endres,J. Hess,J. Sturm,et al.*3-Dmappingwithan RGB-D camera.*IEEE Transactions on Robotics,30(1): 17-187 (2017).

[8] M. Bloesch,S. Omari,M. Hutter,et al.*Robust visual inertial odometry using a direct EKF- based approach* In Proceedings of International Conference on Intelligent Robots and Systems, 2015:298-304 (2015).

[9] S. Leutenegger,S. Lynen,M. Bosse,et al.*Keyframe-based visual-inertial odometry using nonlinear optimization.*International Journal of Robotics Research,34(3): 314-334 (2015).

[10] T. Qin,P. Li,S. Shen. *VINS-Mono:A robust and versatile monocular visual-inertial state estimator*.arXiv,2006: 1708. 03852v1 (2017).

[11] H. Strasdat,J M M. Montiel,A J. Davison. *Visual SLAM: Why filter?* Image & Vision Computing,30(2): 65-77 (2012).

[12] H. Strasdat,J M M. Montiel,A J. Davison.*Real-time monocular SLAM:Why filter?* In Proceedings of IEEE International Conference on Robotics and Automation, 2010:2657-2644 (2010).

[13] A. Handa,M. Chli,H. Strasdat,et al.*Scalable active matching* In Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, 2010:1546-1553 (2010).

[14] J. Civera,O G. Grasa,A J. Davison,et al.*1-Point RANSAC for extended Kalman filtering:Application to real-time structure from motion and visual odometry*. Journal of Field Robotics,2010,27(5): 609-631 (2010).

[15] C. Mei,G. Sibley,M. Cummins,et al.*RSLAM:A system for large-scale mapping in constant-time using stereo*.International Journal of Computer Vision,2011, 94(2): 198-214 (2011).

[16] R. Mur-Artal,J D. Tardos.*ORB-SLAM2:An open-source SLAM system for monocular,stereo,and RGB-D cameras*.IEEE Transactions on Robotics,2016,33(5): 1255-1262 (2016).

[17] Xiang Gao, Tao Zhang, Yi Liu, Qinrui Yan.*14 Lectures on Visual SLAM: From Theory to Practice*, Publishing House of Electronics Industry (2017).

[18] A J. Davison,I D. Reid,N D. Molton,et al.*MonoSLAM: Real-time single camera SLAM*. IEEE Transactions on Pattern Analysis & Machine Intelligence,29(6): 1052 (2007).

[19] G. Klein,D. Murray.*Parallel tracking and mapping for small AR workspaces* In Proc of IEEE & ACM Int Sympo on Mixed & Augmented Reality,2007: 1-10 (2007).

[20] C. Forster,M. Pizzoli,D. Scaramuzza.*SVO:Fast semi-direct monocular visual odometry* In Proceedings of IEEE International Conference on Robotics and Automation,2014:15-22 (2014).

[21] J. Engel,V. Koltun,D. Cremers.*Direct sparse odometry*.IEEE Transactions on Pattern Analysis & Machine Intelligence,40(3): 611-625 (2017).

[22] J. Engel,T. Schops,D. Cremers.*LSD-SLAM:Large-scale direct monocular SLAM* In PProceedings of European Conference on Computer Vision, 2014:834-849 (2014).

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

[24] M. Labbe,F. Michaud.*Online global loop closure detection for large-scale multi-session graph-based SLAM* In Proceedings of International Conference on Intelligent Robots and Systems,2014:2661-2666 (2014).

[25] R A. Newcombe,S J. Lovegrove,A J. Davison.*DTAM: Dense tracking and mapping in real-time* In Proceedings of IEEE International Conference on Computer Vision,2011:2320-2327 (2011).

[26] C. Kerl,J. Sturm,D. Cremers.*Dense visual SLAM for RGB-D cameras* In Proceedings of International Conference on Intelligent Robots and Systems,2014:2100-2106 (2014).

[27] T. Whelan,R F. Salas-Moreno,B. Glocker,et al.*ElasticFusion:Real-time dense SLAM and light source estimation*.International Journal of Robotics Research,2016,35(14): 1697-1716 (2016).

[28] T. Whelan,S. Leutenegger,R S. Moreno,et al.*ElasticFusion:Dense SLAM without a pose graph*.International Journal of Robotics Research,2016,35(14): 1-9 (2016).

[29] M. Bloesch,M. Burri,S. Omari,et al.*Iterated extended Kalman filter based visual-inertial odometry using direct photometric feedback*.International Journal of Robotics Research,36(10): 1053-1072 (2017).

[30] S. Izadi,D. Kim,O. Hilliges,et al.*KinectFusion:Real-time 3D reconstruction and interaction using a moving depth camera* In Proceedings of ACM Symposium on User Interface Software and Technology,Santa Barbara,CA, USA,559-568 (2011).

[31] K. Konolige,M. Agrawal,J. Sola.*Large-scale visual odometry for rough terrain* In Proceedings of International Symposium on Robotics Research,November 26-29,201-212 (2011).

[32] S D. Quijada,E. Zalama,J G. Garcia-Bermejo,et al.*Fast 6D odometry based on visual features and depth* In Intelligent Autonomous Systems 12.Berlin Heidelberg: Springer,2013:5-26 (2013).

[33] C. Tang,O. Wang,P. Tan.*GlobalSLAM:Initialization-robust Monocular Visual SLAM*. arXiv:1708.04814v1 (2017).

[34] D. Scaramuzza,F. Fraundorfer.*Visual Odometry[Tutorial]*.IEEE Robotics & Automation Magazine,18(4): 80-92 (2011).

[35] R I. Hartley.*In defense of the eight-point algorithm*.IEEE Transactions on Pattern Analysis & Machine Intelligence,19(6): 580-593 (1997).

[36] P J. Besl,N D. Mckay.*A method for registration of 3-Dshapes*.IEEE Transactions on Pattern Analysis and Machine Intelligence,14(2): 239-256 (1992).

[37] A. Penate-Sanchez,J. Andrade-Cetto,F. Moreno-Noguer.*Exhaustive linearization for robust camera pose and focal length estimation*.IEEE Transactions on Pattern Analysis and Machine Intelligence,35(10): 2387-2400 (2013).

[38] V. Lepetit,F. Moreno-Noguer,P. Fua.*EPnP:An accurate O(n) solution to the PnP problem*. international Journal of Computer Vision,81(2): 155-166 (2009).

[39] Wikipedia.*Feature(computer vision)* In $https://enwikipediaorg/wiki/Feature_(computer_vision)$, (2016-07-09), [2017-11-01].

[40] D G. Lowe.*Distinctive Image features from scale-invariant key points*.International Journal of Computer Vision, 60(2): 155-166 (2004).

[41] H. Bay,T. Tuytelaars,L V. Gool.*SURF:Speeded up robust features* In Proceedings of European Conference on Computer Vision, 2006:404-417 (2006).

[42] E. Rublee,V. Rabaud,K. Konolige,et al.*ORB:An efficient alternative to SIFT or SURF* In Proceedings of IEEE International Conference on Computer Vision, 2012:2564-2571 (2012).

[43] S. Leutenegger,M. Chli,R Y. Siegwart.*BRISK:Binary robust invariant scalable keypoints* In Proceedings of International Conference on Computer Vision, 2011: 2548-2555 (2011).

[44] E. Rosten,T. Drummond.*Machine learning for high-speed corner detection* In Proceedings of European Conference on Computer Vision, 2006: 430-443 (2006).

[45] M. Calonder,V. Lepetit,C. Strecha,et al.*BRIEF:Binary robust independent elementary feature* In Proceedings of European Conference on Computer Vision, 2010: 778-792 (2010).

[46] B. Kitt,A. Geiger,H. Lategahn.*Visual odometry based on stereo image sequences with RANSAC-based outlier*

rejection scheme In Proceedings of Intelligent Vehicles Symposium, 2010:486-492 (2010).

[47] A. Geiger,J. Ziegler,C. Stiller.*StereoScan:Dense 3D reconstruction in real-time* In Proceedings of IEEE Intelligent Vehicles Symposium, 2011:963-968 (2011).

[48] S. Baker,I. Matthews.*Lucas-Kanade 20 years on:A unifying framework*. International Journal of Computer Vision,2004,56(3): 221-255 (2004).

[49] P. Favaro,H. Jin,S. Soatto.*A semi-direct approach to structure from motion* In Proceedings of International Conference on Image Analysis and Processing, 2001:250-255 (2001).

[50] S. Benhimane,E. Malis.*Integration of Euclidean constraints in template based visual tracking of piecewise-planar scenes* In Proceedings of International Conference on Intelligent Robots and Systems,2007:1218-1223 (2007).

[51] G. Silveira,E. Malis,P. Rives.*An efficient direct approach to visual SLAM*. IEEE Transactions on Robotics,2008,24(5): 969-979 (2008).

[52] T. Gokhool,M. Meilland,.P Rives,et al.*A dense map building approach from spherical RGBD images* In Proceedings of International Conference on Computer Vision Theory and Applications,2014:656-663 (2014).

[53] M. Pizzoli,C. Forster,D. Scaramuzza.*REMODE:Probabilistic,monocular dense reconstruction in real time* In Proceedings of IEEE International Conference on RoboticsandAutomation,2014:2609-2616 (2014).

[54] J. Engel,D. Cremers.*Semi-dense visual odometry for a monocular camera* In Proceedings of IEEE International Conference on Computer Vision,2014:1449-1456 (2014).

[55] G. Vogiatzis,C. Hernandez.*Video-based,real-time multi-view stereo*.Image & Vision Computing,29(7): 434-441 (2011).

[56] C. Forster,Z. Zhang,M. Gassner,et al.*SVO:Semidirect visual odometry for monocular and multicamera systems*.IEEE Transactions on Robotics,33(2): 249-265 (2017).

[57] S. Lovegrove,A J. Davison,J. Ibanez-Guzman.*Accurate visual odometry from a rear parking camera* In Proceedings of Intelligent Vehicles Symposium,2011:788-793 (2011).

[58] A Z. Zhu,N. Atanasov,K. Daniilidis.*Event-based visual inertial odometry* In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition,2017:5816-5824 (2017).

[59] Y. Lin,F. Gao,T. Qin,et al.*Autonomous aerial navigation using monocular visual-inertial fusion*.Journal of Field Robotics,35(4): 23-51 (2018).

[60] J. Gui,D. Gu,S. Wang,et al.*A review of visual inertial odometry from filtering and optimization perspectives*. Advanced Robotics,29(20): 1289-1301 (2015).

[61] S. Weiss,M W. Achtelik,S. Lynen,et al.*Real-time onboard visual-inertial state estimation and self-calibration of MAVs in unknown environments* In Proceedings of IEEE International Conference on Robotics and Automation, 2012:957-964 (2012).

[62] F. Dellaert,A. Ranganathan,M. Kaess.*Fast 3D poseestimation with out-of-sequence measurements* In Proceedings of IEEE International Conference on Intelligent Robots and Systems, 2007:2486-2493 (2007).

[63] V. Indelman,S. Williams,M. Kaess, et al.*Factor graph based incremental smoothing in inertial navigation systems* In International Conference on Information Fusion, IEEE, 2012:2154-2161 (2012).

[64] S. Yang,S A. Scherer,X. Yi, et al.*Multi-camera visual SLAM for autonomous navigation of micro aerial vehicles*. Robotics & Autonomous Systems,93 (2017).

[65] V. Usenko,J. Engel,J. Stuckler, et al.*Direct visual-inertial odometry with stereo cameras* In IEEE International Conference on Robotics and Automation, IEEE, 2016:1885-1892 (2016).

[66] AP. Shetty.*GPS-LiDAR sensor fusion aided by 3D city models for UAVs* (2017).

[67] A. Zeng,S. Song,M. Niebner, et al.*3DMatch: Learning Local Geometric Descriptors from RGB-D Reconstructions*.2017:199-208 (2017).

[68] A. Shaked,L. Wolf. Improved Stereo Matching with Constant Highway Networks and Reflective Confidence Learning.2016:6901-6910 (2016).

[69] K. Tateno,F. Tombari,I. Laina, et al.*CNN-SLAM: RealTime Dense Monocular SLAM with Learned Depth Prediction*.2017:6565-6574 (2017).

[70] P. Muller,A. Savakis.*Flowdometry: An Optical Flow and Deep Learning Based Approach to Visual Odometry* In Applications of Computer Vision,IEEE, 2017:624-631 (2017).

[71] X. Gao,T. Zhang.*Robust RGB-D simultaneous localization and mapping using planar point feature*.Robotics & Autonomous Systems,72:1-14 (2015).

[72] S. Yang,S. Scherer.*Direct monocular odometry using points and lines* In IEEE International Conference on Robotics and Automation,IEEE,2017:3871-3877 (2017).

[73] T. Wang,H. Ling.*Gracker: A Graph-based Planar Object Tracker*.IEEE Transactions on Pattern Analysis & Machine Intelligence, PP(99):1-1 (2017).

[74] T. Schops, T. Sattler, C. Hane, et al.*3D Modeling on the Go: Interactive 3D Reconstruction of Large-Scale Scenes on Mobile Devices* In International Conference on 3d Vision, IEEE Computer Society, 2015:291-299 (2015).