The Visual Computer (2020) 36:1053–1065
https://doi.org/10.1007/s00371-019-01714-6

SURVEY

A review of monocular visual odometry

Ming He[1] · Chaozheng Zhu[1] · Qian Huang[2,3] · Baosen Ren[4] · Jintao Liu[1]

Published online: 25 June 2019

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 odometry and basic odometry. This paper describes the problem of visual odometry and also determines the relationships between visual odometry and visual simultaneous localization and mapping (SLAM). The basic principle of visual odometry is expressed in the form of mathematics, specifically by incrementally solving the pose changes of two series of 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 operations, which are generally of interest in the current visual odometry research, are discussed from the future development of the directions and trends. Furthermore, we present a novel framework for the implementation of next-generation visual odometry based on additional high-dimensional features, which have not been implemented in the relevant applications.

Keywords Visual odometry · Multi-sensor data fusion · Machine learning · Visual SLAM

1 Introduction

Due to the complexity of the unknown environment, it is of great significance to build a real-time map and localization based only on the robot's own sensor [1,2]. Visual sensor, which is a common type of robot sensor, has the advantages of high accuracy, low cost and abundant data information. Therefore, using a visual sensor to determine its location has become a main topic of research. The concept of visual odometry [3], proposed by Nister, i.e., correlation image sequences, is analyzed to estimate the mobile robot pose (e.g., position and attitude) in real time through machine vision technology. This process can also overcome the shortcomings of traditional odometries and provide more accurate positioning. Furthermore, it can run where the global positioning system (GPS) is not available, such as indoor environments or interplanetary exploration [3,4].

Visual odometry (VO) was known by the public, when it had been successfully applied to the Mars exploration [4]. It also highlights its important application value in the fields of public security, virtual reality (VR) [5], augmented reality (AR) [6] and so on, as shown in Fig. 1. We have added many latest contents since 2017 based on document [4]. In particular, some novel views on the method of visual state estimation are presented.

✉ Chaozheng Zhu
370045744@qq.com

Ming He
1091721005@qq.com

Qian Huang
huangqian@hhu.edu.cn

Baosen Ren
18354261031@163.com

Jintao Liu
top2012@163.com

1 College of Command and Control Engineering, Army Engineering University of PLA, Nanjing, China

2 College of Computer and Information, HoHai University, Nanjing, China

3 Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun, China

4 State Grid Shandong Electric Power Maintenance Company, Linyi, China

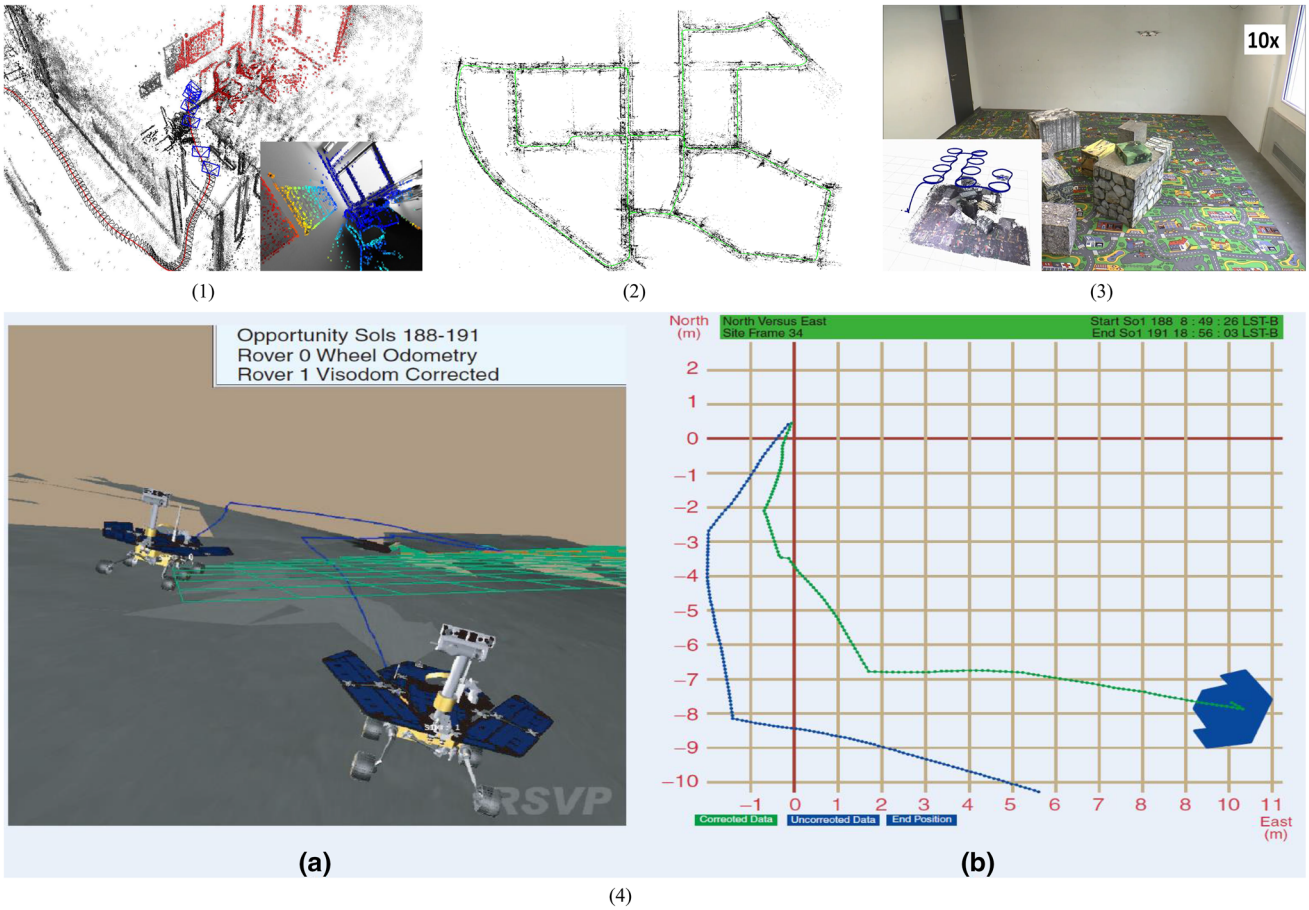


Fig. 1 Research and application of VO samples, where (1), (2) and (3) represent the state-of-the-art research on VO in the form of DSO, ORB-SLAM2 and SVO, respectively. (4) is a real application in 2004 that illustrates: the use of visual odometry on Mars

2 Research problem

The main research problem of VO is how to estimate the trajectory of the camera according to images. For visual simultaneous localization and mapping (vSLAM), there is a major research issue, that is, to generate a loop and effectively integrate new constraints to the current map based on VO.

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

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 a further research on VO. In terms of application scenarios, VO is sufficient in such situations where real-time localization is needed, such as missile guidance flights, unmanned aerial vehicle and AR. However, this aspect is redundant for vSLAM to build an accurate map, which could waste additional computing power.

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. The target of VO is an incremental reconstruction trajectory, which may only optimize the pose of the previous paths. Therefore, VO is called sliding-window-based bundle adjustment. The sliding-window-based optimization relies on a local map in vSLAM.

In recent years, significant progresses [13–16] have been made in both monocular and binocular cameras. Most of these devices can operate in a 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 extended Kalman filtering to optimization [18,19].

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 [20]	2011	Monocular	Direct method	Optimization	Direct method, monocular dense map, needs GPU support
Kinect fusion [21]	2011	RGB-D	Direct method	Optimization	First implementation of dense reconstruction based on RGB-D in real time
DVO [22]	2013	RGB-D	Direct method	Optimization	Direct method based on RGB-D, dense map
SVO [23]	2014	Monocular	Semi-direct method	Optimization	Sparse direct method, only VO
LSD-SLAM [22]	2014	Monocular	Direct method	Optimization	Direct method + semi-dense map
OKVIS [9]	2015	Monocular/multi-cameras + inertial measurement unit (IMU)	Feature point method	Optimization	Mainly optimization based on key frame VIO
ROVIO [8,25]	2015	Monocular + IMU	Direct method	EKF Filter	Mainly EKF based on VIO
Elastic fusion [26,27]	2015	RGB-D	Direct method	Optimization	RGB-D reconstruction in real time, visualization
DSO [28]	2016	Monocular	Direct method	Optimization	Monocular direct method, best results of the direct method at present
ORB-SLAM2 [16,29]	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
Maplab [30]	2018	Monocular/multi-cameras + IMU	Direct method	EKF filter/optimization	Visual-inertial mapping and localization system/the research community with a collection of multi-session mapping tools

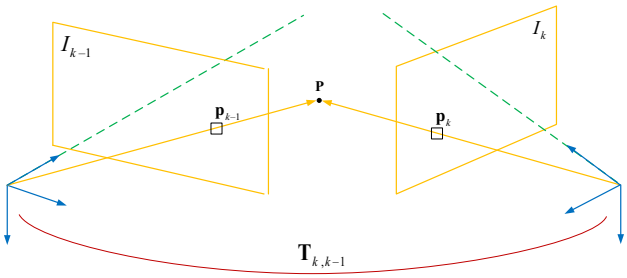


Fig. 2 Illustration of the VO problem

3 Formal description

For a monocular VO, at time k, the image sets designated $I_{0:k} = I_0, \ldots, I_k$, are collected by the camera of a rigid robot. Suppose that the camera coordinates are those of the robot. Nevertheless, in stereo vision systems, the left camera is typically the original.

However, the use of binocular VO leads to a sharp decline in the accuracy of triangulation, as the distance between the center of the two cameras is affected by the conditions of the measurement accuracy and climate changes (e.g., thermal expansion and cool contraction). Therefore, this paper focuses on the research of the monocular VO problem, as shown in Fig. 2.

A rigid transformation $T_{k,k-1} \in \mathbb{R}^{4\times4}$ is formed by two neighbor camera poses from time $k-1$ and k, which is shown 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 $T_{1:k} = \{T_{1,0}, T_{2,1}, \ldots, T_{k,k-1}\}$ contains all the motion sequences. Finally, the camera pose set is $C_{0:k} = \{C_0, C_1, \ldots, C_k\}$, where C_k is the initial coordinates at time k. The current pose C_n can be calculated by the connection between the transformation $T_k (k = 1, 2, \ldots, n)$. In general, $C_n = C_{n-1}T_n$. And C_0 is the camera pose at time $k = 0$.

The main goal of VO is to calculate T_k from image I_k to image I_{k-1} and then integrate all the transformations to restore the entire path $C_{0:k}$ of the camera. In this paper, VO is an incremental reconstruction trajectory. An iterative optimization based on the previous 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 previous m frames (e.g., sliding-window-based bundle adjustment, since it executes on a 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 and t, so that for all feature points z^j, the cumulative error of the two norms is minimal, and the results are as follows:

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

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 adjustment. The principle of bundle adjustment and optimization is shown in Fig. 3.

4 Research status

Significant progress has been made in the VO research [24,31] of large-scale scenes. The methods of VO implementation include the feature points method, the direct tracking method and the hybrid semi-direct tracking method.

4.1 Method based on feature points

For feature point-based method [3,6,16,18,29,32,33], 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 that for each new image I_k, while it is 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, respectively. 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. Actually, there is an assumption that the camera has been calibrated for the majority of VO implementations. The third step is to calculate the relative motion T_k between $k-1$ and k. The choice of method should base on the prior information. For instance, a typical 2D–2D problem incurs when the monocular camera is used while additional information is missing, then it could be solved by the epipolar geometry. Meanwhile, perspective-three-point (P3P) and iterative closet point (ICP) [35–38] are suitable for the other cases. The pose of the camera C_k is based on the transformation of the previous pose T_k. To achieve more precise local trajectory estimation by iterative optimization (e.g., bundle adjustment), we build a local map based on the depth estimation of the previous m frames.

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 decentral-

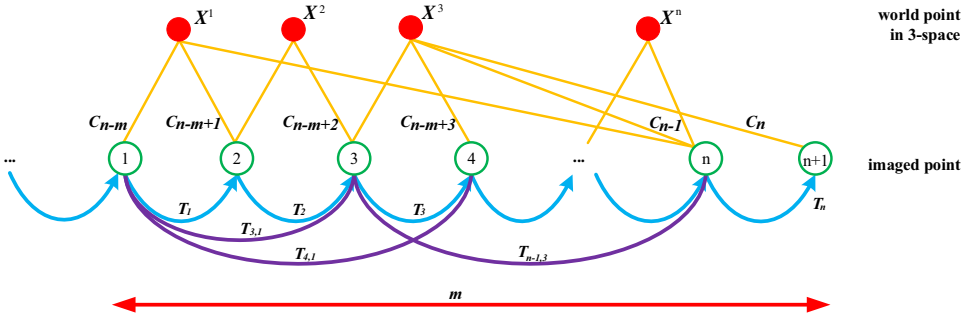


Fig. 3 Principle of bundle adjustment and optimization, where C represents the camera pose of the current frame, T represents the transformation of the pose between the two cameras and m represents the total number of cameras

ized nature of the outliers, the random sampling consistency (RANSAC) is used to select the optimal matching, but not the least-square matching algorithm.

Typically, due to the influence of illumination and deformation on the gray value, the change between different images may be considerable. Therefore, mere 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 neighborhood 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 defined as feature.

With full consideration to all kinds of problems on the process of image transformation, speeded-up robust features (SURF) [40] and scale-invariant feature transform (SIFT) [41] still cost a lot of computation. Generally, it could be challenging to execute calculation in real time on a CPU. However, the popularity of some computable feature extraction/description algorithms, such as oriented FAST and rotated BRIEF (ORB) [42] and Binary Robust Invariant Scalable Keypoints (BRISK) [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 Features from Accelerated Segment Test (FAST) [44] and Binary Robust Independent Elementary Features (BRIEF), 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 consist of key points and descriptors. Among them, for corner extraction, it increases the main direction of the feature points on the basis of FAST to add rotation invariance in the descriptor of ORB. Additionally, for the new BRIEF descriptor: it is a method to describe the pixel area which surrounds the key points extracted earlier. As 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.

Table 2 Performance comparison between different features

		Feature type		
		ORB	SURF	SIFT
Complexity		√√√√√	√√√	√
Robustness	Rotation, blur	√	√√√	√
	Scale-variance	×	√√√	√

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.

Early real-time VO was based on the feature point. For example, the monocular VO framework (e.g., 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 localization 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 on the basis of the optical flow and deep learning [46]. 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 the existing research is ORB-SLAM2 [16], which presents a more complete VO framework. 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 pose 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. Besides, this scheme is also compatible with monocular, binocular and RGB-D cameras.

For initialization, [16] proposes an strategy for automatic initialization map and calculates the homography matrix (i.e., assuming a planar scene) [31] and essential matrix (i.e., assuming non-planar scene) [32]. According to the heuristic rule, the corresponding situation is determined to initialize the pose. This is also a remarkable 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.

4.2 Method based on direct tracking

The direct method of estimating camera motion based on the pixel gray invariance hypothesis has developed rapidly in recent years [20,28]. The direct method, which is developed from the optical flow [47], can estimate the camera motion and the pixel's spatial location by minimizing the photometric error (i.e., minimizing the reprojection error of feature points in the feature point method) without extracting the feature or 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 [48–50]. Recently, it appears that the direct methods could use directly 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 pose graph [28,51]. To construct a semi-dense 3D environment map, Engel et al. [24] 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 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 graph, so it can form a closed loop and accommodate large-scale scenarios. The system selects the nearest key frame for each newly added key frame in the existing key frame set (i.e., map).

Direct sparse odometry (DSO) [28] 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. As the new depth estimation mechanism is used to optimize the sliding window instead of the original Kalman filtering method, it provides an improvement in accuracy. In addition, in contrast to LSD-SLAM, DTAM [20] provides a direct method to calculate a real-time dense map based on a monocular camera. The pose 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 [52,53].

4.3 Method based on the hybrid semi-direct tracking

Based on the advantages of the feature-based method and the direct tracking method, a hybrid semi-direct method is proposed, namely semi-direct visual odometry (SVO) [23]. Although SVO is still dependent on the characteristics of consistency, this method applies the direct method to obtain the pose. 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 (i.e., ARM Cortex A9 1.6 GHz CPU), and the frame rate can be as high as 300 fps on a general laptop (i.e., Intel i7 2.8 GHz CPU).

Depth estimation is the core of building a local point cloud map. In terms of depth estimation, SVO is 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 [54], while LSD-SLAM is based on the Gauss distribution model. First, the direct method is used to solve the pose matching. Second, classical Lucas–Kanade optical flow [47] matching is used to obtain subpixel accuracy. Then, the minimized reprojec-

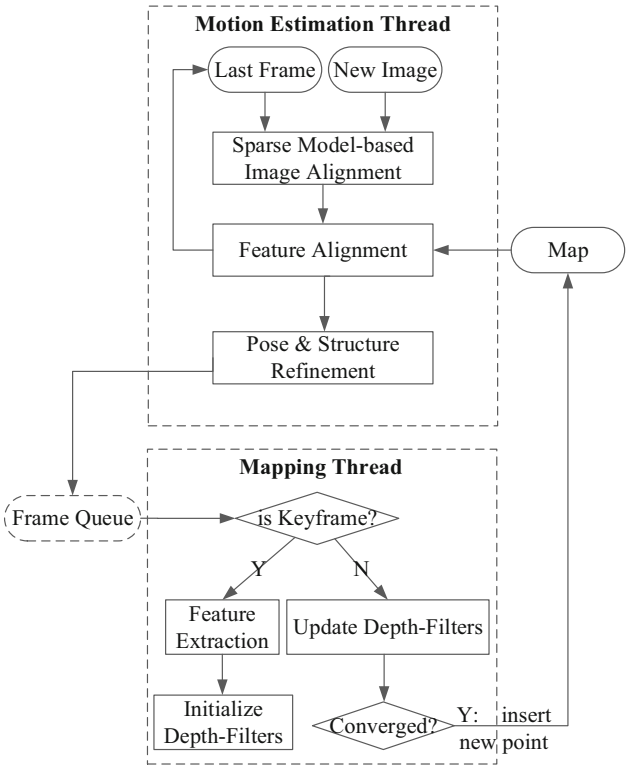


Fig. 4 Module flow chart of SVO

tion error is optimized by combining the point cloud map, as shown in Fig. 4.

In contrast to using the traditional feature points, the whole process needs to rely on feature points when selecting key frames only. The calculation of matching descriptors is erased, and the steps of using RANSAC to remove the outliers are effaced, so the process is relatively efficient; relative to the direct method, this method does not directly match the whole image to obtain the pose of the camera. Instead, it extracts the image block from the whole image, allowing us to obtain the pose from the image block. This technique enhances the robustness of the algorithm. The largest contribution of SVO is the design of the three optimization methods (i.e., optimize the gray error, the feature point prediction position and 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 multi-camera systems [55], tracking the edge, including the prior knowledge of motion. The method also supports various cameras, such as fish-eye and perspective cameras.

4.4 Analysis

The feature point method has been widely used, but its robustness is mainly based on the description of feature points. On the one hand, as robustness enhances, the complexity of the feature point description is increased, which leads to a large increase in the complexity of the algorithm. On the other hand, the feature point method cannot be applied to scenes with weaker feature points, such as wall and sky.

VO [16,29] based on feature points is more mainstream. However, from the experimental results by TUM group of the University of Munich, the direct method of VO [23,24,28] has made great breakthroughs in recent years. Among them, the sparse direct method [28] has a faster and better performance when compared with 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 [28], the direct tracking method is more sensitive to geometric noise, for instance, that produced by a rolling shutter camera. Feature-based methods are more sensitive to optical noise, such as fuzzy noise. Therefore, for common mobile devices (e.g., the shutter camera), the feature-based method might have a better performance. For a robot equipped with the global shutter camera, the method based on direct tracking is becoming increasingly widely used.

The direct method is a relatively new method that can be adapted to scenes with insufficient features, such as corridors or smooth walls [56], 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 non-convexity, single-pixel non-segmentation and the poorly supported assumption of gray invariance in the direct method; thus, its research and implementation are not as stable as that of the feature point method. At present, the direct methods are suitable only for the situations of small motion and brightness changes.

Although methods based on direct tracking are popular, a low speed and lack of assurance of optimality or consistency are problems of the direct method. Therefore, a method based on hybrid semi-direct tracking [23], 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 experimental video and opened 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.

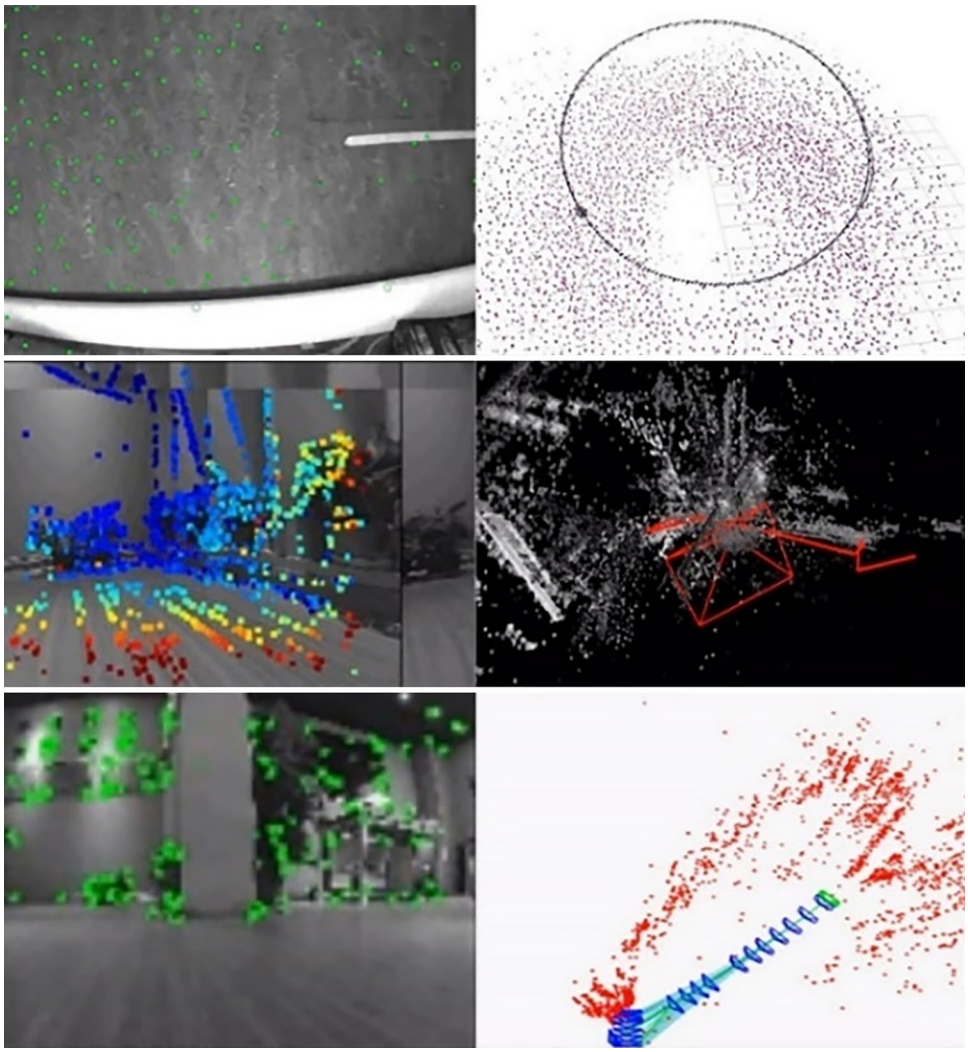


Fig. 5 Comparison of the results of the three types of VO methods [4]. Top line: SVO algorithm; middle line: DSO algorithm; bottom line: ORB-SLAM2 algorithm

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 Fig. 5. We can easily find that the feature method and the hybrid semi-direct method can only build a sparse map; nevertheless, the direct method can build a semi-dense map.

5 Development trends and active research areas

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

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 new sensors exploration, multi-sensor data fusion, machine learning-based research, high-dimensional information mining and a novel framework of VO.

5.1 New sensors exploration

Microsoft's RGB-D camera Kinect, which was released in 2010, can obtain a depth map in real time and simplify calculations substantially, enabling the realization of dense 3D reconstruction systems [7,21,22,26,27]. 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 et al. of the University of

Table 3 Research direction of frontier institutions

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 visual-inertial system
MI	Laser vision-multiple sensors fusion

Zurich proposed a VO algorithm based on the event camera. Moreover, based on the EKF and the unstructured measurement model, IMU was integrated as a complement to data fusion to accurately obtain the pose of University of Zurich proposed VO algorithm based on event camera. Additionally, based on the EKF and the unstructured measurement model, IMU is integrated as a complement to data fusion to get the pose [57] of a 6-DOF camera.

5.2 Multi-sensor 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 [8–10,31,58] has also been a notable trend in recent years. Apple Inc's ARKit, released 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 integrate multiocular and inertial navigation data with the optimized key frame [59]. 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 [60] used visual pose estimation to maintain an EKF of an indirect IMU. Similar loose coupling algorithms include [61,62]; here, the pose 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 (i.e., VINS), so the high-precision visual-inertial navigation system is tightly coupled. Many researchers have explored multi-sensor fusion, e.g., the integration of multi-camera sensors [63] proposed by Yang Shaowu with binocular stereo vision and inertial navigation, speed and data fusion [64]. 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 [65].

5.3 Machine learning-based researches

In recent years, machine learning methods such as neural networks have caused a widespread academic sensation in many fields, and the VO field is of no exception. In the matching tracking part, a data-driven model (i.e., 3DMatch) was proposed [66]. 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 [67] on this new architecture using the stereo matching benchmark dataset showed that the matching reliability is far superior to the existing algorithms.

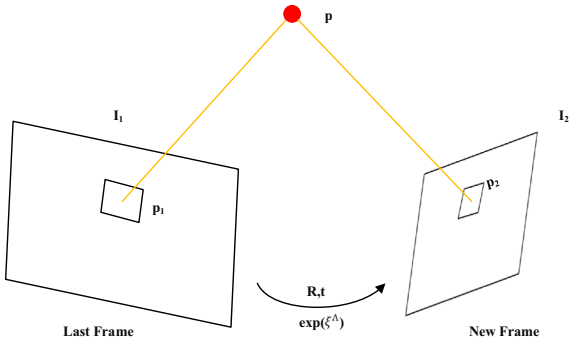


Fig. 6 Direct high-dimensional optical flow diagram

The lack of scale information in monocular VO has been the issue of greatest concern for researchers. Recently, German researchers, such as Keisuke et al., addressed the failure of recovering scale for monocular VO such as low-texture areas, who proposed a fusion method include depth information predicted by 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 [68].

In 2018, Gomez Ojeda et al. [69] proposed a method called learning-based image enhancement for VO in Challenging High Dynamic Range (HDR) Environments. They also proposed a convolutional neural network of reduced size capable of performing faster and overcome one of the main open challenges in VO which is the robustness to difficult illumination conditions or HDR environments. Besides, in terms of pose accuracy, by applying deep recurrent convolutional neural networks (RCNNs), a novel monocular VO system call UnDeepVO [70] and a novel end-to-end framework for monocular VO are confirmed to have a better performance than other monocular VO methods. Therefore, the application of machine leaning shall become the next hot spot in the VO field.

5.4 High-dimensional information mining

The dependency of VO on scene features is essentially due to the use of the overly underlying local features (i.e., 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 (i.e., more regular items) are expected to provide better robustness. For instance, in aspect of edge feature, Yang et al. [72] proposed a monocular VO algorithm that combined point and edge advantages. This algorithm not only performed well in the monocular open dataset [28] which was provided by TUM, but also greatly reduced the motion estimation error in low-texture environments. Li et al. [73] proposed an extension to a point-based direct monocular visual odometry method. It used lines to guide keypoint selection rather than acting as features. Thereby, it can augment efficiency and accuracy. At the same time, in aspect of planar feature, Wang et al. [74] 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.

5.5 A novel framework of VO

The ORB feature contains 4-DoF information including scale invariance (z), rotation invariance (θ) and translation invariance (x, y). In contrast to previous classical LK optical flow, in which only 2-DoF of a camera can be obtained [47], it is possible to provide a new 1-DoF to a camera by the improved corner feature (e.g., ORB). 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 6-DoF problem for a camera, including rotation and translation and then incrementally solve the VO problem, as shown in Fig. 6. By eliminating the need for costly feature matching and decomposing the essential matrix (i.e., as required by conventional methods), it 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.

6 Conclusion

This paper analyzes the differences between VO and vSLAM and formalizes the VO problem. Then, we focus on the status of various methods to implement VO. After that, their features are compared by a series of tests. At present, most researchers focus mainly on ideal scenes with a satisfactory visual field, such as daytime. However, a variety of both indoor and outdoor scenes (i.e., from day to night and with seasons changing) are very common. How to ensure system robustness under such circumstances is an important research direction. Besides, we present the concept of similar optical flow method in the last part of this paper, which may decrease the complexity of VO. In the future, we will focus on the application of a novel framework of VO. in particular, in certain harsh environment such as large-scale indoor fire, helping firefighters to be positioned and draw motion trajec-

tory in real time is a vital method to improve the efficiency of search-and-rescue work.

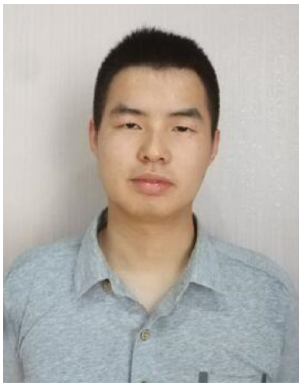
Acknowledgements This work was supported by National Key R&D Program of China Nos. 2018YFC0806900, 2016YFC0800606, 2016YFC0800310 and 2018YFC0407905; Natural Science Foundation of Jiangsu Province under Grants No. BK20161469; Primary Research & Development Plan of Jiangsu Province under Grant Nos. BE2016904, BE2017616, and BE2018754.

References

1. Durrant-Whyte, H., Bailey, T.: Simultaneous localization and mapping: part I. IEEE Robot. Autom. Mag. **13**(2), 99–110 (2006)
2. Durrant-Whyte, H., Bailey, T.: Simultaneous localization and mapping: part II. IEEE Robot. Autom. Mag. **13**(2), 99–110 (2006)
3. Nistér, D., Naroditsky, O., Bergen, J.: Visual odometry. In: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004 (CVPR 2004), vol. 1, pp. I–I. IEEE (2004)
4. Zhu, C., He, M., et al.: A survey of monocular visual odometry. Comput. Eng. Appl. **54**(07), 20–28+55 (2018). **(in Chinese with English abstract)**
5. Lin, S., Chen, Y., Lai, Y.K., Martin, R.R., Cheng, Z.Q.: Fast capture of textured full-body avatar with rgb-d cameras. Vis. Comput. **32**(6–8), 681–691 (2016)
6. Sharma, O., Pandey, J., Akhtar, H., Rathee, G.: Navigation in AR based on digital replicas. Vis. Comput. **34**(6–8), 925–936 (2018)
7. Teng, C.H., Chuo, K.Y., Hsieh, C.Y.: Reconstructing three-dimensional models of objects using a kinect sensor. Vis. Comput. **34**, 1507–1523 (2018)
8. Bloesch, M., Omari, S., Hutter, M.,Siegwart, R..: Robust visual inertial odometry using a direct EKF-based approach. In: 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 298–304. IEEE (2015)
9. Kai, W., Liwei, L., Yong, L., Peng, D., Guoting, X.: Application research of chaotic carrier frequency modulation technology in two-stage matrix converter. Math. Probl. Eng. **2019**(2614327), 8 (2019). https://doi.org/10.1155/2019/2614327
10. Qin, T., Li, P., Shen, S.: Vins-mono: a robust and versatile monocular visual-inertial state estimator. IEEE Trans. Robot. **34**(4), 1004–1020 (2018)
11. Strasdat, H., Montiel, J.M., Davison, A.J.: Visual SLAM: Why filter? Image Vis. Comput. **30**(2), 65–77 (2012)
12. Strasdat, H., Montiel, J.M.M., Davison, A.J.: Real-time monocular SLAM: Why filter?. In: 2010 IEEE International Conference on Robotics and Automation (ICRA), pp. 2657–2664. IEEE (2010)
13. Kai, W., JinBo, P., LiWei, L., Shengzhe, Z., Yuhao, L., Tiezhu, Z.: Synthesis of hydrophobic carbon nanotubes/reduced graphene oxide composite films by flash light irradiation. Front. Chem. Sci. Eng. **12**(3), 376–382 (2018)
14. Kai, W., ShengZhe, Z., YanTing, Z., Jun, R., LiWei, L., Yong, L.: Synthesis of porous carbon by activation method and its electrochemical performance. Int. J. Electrochem. Sci. **13**(11), 10766–10773 (2018)
15. Mei, C., Sibley, G., Cummins, M., et al.: RSLAM: a system for large-scale mapping in constant-time using stereo. Int. J. Comput. Vision **94**(2), 198–214 (2011)
16. MMur-Artal, R., Tardós, J.D.: Orb-slam2: an open-source slam system for monocular, stereo, and rgb-d cameras. IEEE Trans. Robot. **33**(5), 1255–1262 (2017)
17. Gao, X., Zhang, T., Liu, Y., Yan, Q.: Lectures on Visual SLAM: From Theory to Practice. Publishing House of Electronics Industry, Beijing (2017)
18. Davison, A.J., Reid, I.D., Molton, N.D., Stasse, O.: MonoSLAM: Real-time single camera SLAM. IEEE Trans. Pattern Anal. Mach. Intell. **6**, 1052–1067 (2007)
19. Klein, G., Murray, D.: Parallel tracking and mapping for small AR workspaces. In: 6th IEEE and ACM International Symposium on Mixed and Augmented Reality, 2007 (ISMAR 2007), pp. 225–234. IEEE (2007)
20. Newcombe, R.A., Lovegrove, S.J., Davison, A.J.: DTAM: dense tracking and mapping in real-time. In: 2011 IEEE International Conference on Computer Vision (ICCV), pp. 2320–2327. IEEE (2011)
21. Izadi, S., Kim, D., Hilliges, O., Molyneaux, D., Newcombe, R., Kohli, P., Shotton, J., Hodges, S., Freeman, D., Davison, A., Fitzgibbon, A.: KinectFusion: real-time 3D reconstruction and interaction using a moving depth camera. In: Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology, pp. 559–568. ACM (2011)
22. Kerl, C., Sturm, J., Cremers, D.: Dense visual SLAM for RGB-D cameras. In: 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 2100–2106. IEEE (2013)
23. Forster, C., Pizzoli, M., Scaramuzza, D.: SVO: Fast semi-direct monocular visual odometry. In: 2014 IEEE International Conference on Robotics and Automation (ICRA), pp. 15–22. IEEE (2014)
24. Engel, J., Schöps, T., Cremers, D.: LSD-SLAM: large-scale direct monocular SLAM. In: European Conference on Computer Vision, pp. 834–849. Springer, Cham (2014)
25. Bloesch, M., Burri, M., Omari, S., Hutter, M., Siegwart, R.: Iterated extended Kalman filter based visual-inertial odometry using direct photometric feedback. Int. J. Robot. Res. **36**(10), 1053–1072 (2017)
26. Whelan, T., Salas-Moreno, R.F., Glocker, B., Davison, A.J., Leutenegger, S.: ElasticFusion: real-time dense SLAM and light source estimation. Int. J. Robot. Res. **35**(14), 1697–1716 (2016)
27. Whelan, T., Salas-Moreno, R.F., Glocker, B., Davison, A.J., Leutenegger, S.: ElasticFusion: dense SLAM without a pose graph. Int. J. Robot. Res. **35**(14), 1–9 (2016)
28. Engel, J., Koltun, V., Cremers, D.: Direct sparse odometry. IEEE Trans. Pattern Anal. Mach. Intell. **40**(3), 611–625 (2018)
29. Mur-Artal, R., Montiel, J.M.M., Tardos, J.D.: ORB-SLAM: a versatile and accurate monocular SLAM system. IEEE Trans. Robot. **31**(5), 1147–1163 (2015)
30. Schneider, T., Dymczyk, M., Fehr, M., Egger, K., Lynen, S., Gilitschenski, I., et al.: Maplab: an open framework for research in visual-inertial mapping and localization. IEEE Robot. Autom. Lett. **3**(3), 1418–1425 (2018)
31. Konolige, K., Agrawal, M., Sola, J.: Large-scale visual odometry for rough terrain. In: Kaneko, M., Nakamura, Y. (eds.) Robotics Research, pp. 201–212. Springer, Berlin (2010)
32. Quijada, S.D., Zalama, E., Garcí-Bermejo, J.G., Worst, R., Behnke, S.: Fast 6D odometry based on visual features and depth. In: Lee, S., Cho, H., Yoon, K.J., Lee, J. (eds.) Intelligent Autonomous Systems 12, pp. 245–256. Springer, Berlin (2013)
33. Tang, C., Wang, O., Tan, P.: GlobalSLAM: initialization-robust monocular visual SLAM (2017). arXiv preprint arXiv:1708.04814
34. Scaramuzza, D., Fraundorfer, F.: Visual odometry [tutorial]. IEEE Robot. Autom. Mag. **18**(4), 80–92 (2011)
35. Hartley, R.I.: In defense of the eight-point algorithm. IEEE Pami **19**(6), 580–593 (1997)
36. Besl, P.J., McKay, N.D.: Method for registration of 3-D shapes. In: Sensor Fusion IV: Control Paradigms and Data Structures, vol. 1611, pp. 586–607. International Society for Optics and Photonics (1992)

37. Persson, M., Nordberg, K.: Lambda twist: an accurate fast robust perspective three point (P3P) solver. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 318–332 (2018)
38. Altantsetseg, E., Khorloo, O., Konno, K.: Rigid registration of noisy point clouds based on higher-dimensional error metrics. Vis. Comput. **34**(6–8), 1021–1030 (2018)
39. Kang, H.Y., Han, J.: Feature-preserving procedural texture. Vis. Comput. **33**(6–8), 761–768 (2017)
40. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**(2), 91–110 (2004)
41. Bay, H., Tuytelaars, T., Van Gool, L.: Surf: speeded up robust features. In: European Conference on Computer Vision, pp. 404–417. Springer, Berlin (2006)
42. Rublee, E., Rabaud, V., Konolige, K., Bradski, G.: ORB: an efficient alternative to SIFT or SURF. In: 2011 IEEE international conference on computer vision (ICCV), pp. 2564–2571. IEEE (2011)
43. Leutenegger, S., Chli, M., Siegwart, R.Y.: BRISK: binary robust invariant scalable keypoints. In: 2011 IEEE International Conference on Computer Vision (ICCV), pp. 2548–2555. IEEE (2011)
44. Rosten, E., Drummond, T.: Machine learning for high-speed corner detection. In: European Conference on Computer Vision, pp. 430–443. Springer, Berlin (2006)
45. Calonder, M., Lepetit, V., Strecha, C., Fua, P.: BRIEF: binary robust independent elementary features. In: European Conference on Computer Vision (2010)
46. Muller, P., Savakis, A.: Flowdometry: an optical flow and deep learning based approach to visual odometry. In: 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 624–631. IEEE (2017)
47. Baker, S., Matthews, I.: Lucas-Kanade 20 years on: a unifying framework. Int. J. Comput. Vis. **56**(3), 221–255 (2004)
48. Lu, F., Zhou, B., Zhang, Y., Zhao, Q.: Real-time 3d scene reconstruction with dynamically moving object using a single depth camera. Vis. Comput. **34**, 753–763 (2018)
49. Jin, H.L., Favaro, P., Soatto, S.: A semi-direct approach to structure from motion. Vis. Comput. **19**(6), 377–394 (2003)
50. Zhou, Y., Yan, F., Zhou, Z.: Handling pure camera rotation in semi-dense monocular SLAM. Vis. Comput. **35**, 123 (2019). https://doi.org/10.1007/s00371-017-1435-0
51. Silveira, G., Malis, E., Rives, P.: An efficient direct approach to visual slam. IEEE Trans. Robot. **24**(5), 969–979 (2008)
52. Pizzoli, M., Forster, C., Scaramuzza, D.: REMODE: probabilistic, monocular dense reconstruction in real time. In: 2014 IEEE International Conference on Robotics and Automation (ICRA), pp. 2609–2616. IEEE (2014)
53. Engel, J., Sturm, J., Cremers, D.: Semi-dense visual odometry for a monocular camera. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1449–1456 (2013)
54. Vogiatzis, G., Hernandez, C.: Video-based, real-time multi-view stereo. Image Vis. Comput. **29**(7), 434–441 (2011)
55. Forster, C., Zhang, Z., Gassner, M., Werlberger, M., Scaramuzza, D.: Svo: semidirect visual odometry for monocular and multicamera systems. IEEE Trans. Robot. **33**(2), 249–265 (2017)
56. Lu, R., Zhu, F., Wu, Q., Fu, X.: Search inliers based on redundant geometric constraints. Vis. Comput. (2018). https://doi.org/10.1007/s00371-018-1605-8
57. Zhu, A.Z., Atanasov, N., Daniilidis, K..: Event-based visual inertial odometry. In: CVPR, pp. 5816–5824 (2017)
58. Lin, Y., Gao, F., Qin, T., Gao, W., Liu, T., Wu, W., Zhenfei, Y., Shen, S.: Autonomous aerial navigation using monocular visual-inertial fusion. J. Field Robot. **35**(1), 23–51 (2018)
59. Gui, J., Gu, D., Wang, S., Hu, H.: A review of visual inertial odometry from filtering and optimisation perspectives. Adv. Robot. **29**(20), 1289–1301(2015)
60. Weiss, S., Achtelik, M.W., Lynen, S., Chli, M., Siegwart, R.: Real-time onboard visual-inertial state estimation and self-calibration of MAVs in unknown environments. In: 2012 IEEE International Conference on Robotics and Automation (ICRA), pp. 957–964. IEEE (2012)
61. Weiss, S., Achtelik, M.W., Lynen, S., Chli, M., Siegwart, R.: Real-time onboard visual-inertial state estimation and self-calibration of MAVs in unknown environments. In: IEEE International Conference on Robotics & Automation. IEEE (2013)
62. Ranganathan, A., Kaess, M., Dellaert, F.: Fast 3D pose estimation with out-of-sequence measurements. In: IEEE/RSJ International Conference on Intelligent Robots & Systems. IEEE (2007)
63. Yang, S., Scherer, S.A., Yi, X., Zell, A.: Multi-camera visual SLAM for autonomous navigation of micro aerial vehicles. Robot. Auton. Syst. **93**, 116–134 (2017)
64. Usenko, V., Engel, J., Stückler, J., Cremers, D.: Direct visual-inertial odometry with stereo cameras. In: 2016 IEEE International Conference on Robotics and Automation (ICRA), pp. 1885–1892. IEEE (2016)
65. Shetty, A.P.: GPS-LiDAR sensor fusion aided by 3D city models for UAVs (Doctoral dissertation) (2017)
66. Zeng, A., Song, S., Nießner, M., Fisher, M., Xiao, J., Funkhouser, T..: 3dmatch: learning local geometric descriptors from rgb-d reconstructions. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 199–208. IEEE (2017)
67. Shaked, A., Wolf, L.: Improved stereo matching with constant highway networks and reflective confidence learning. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4641–4650 (2017)
68. Tateno, K., Tombari, F., Laina, I., Navab, N.: CNN-SLAM: real-time dense monocular SLAM with learned depth prediction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 2 (2017)
69. Gomez-Ojeda, R., Zhang, Z., Gonzalez-Jimenez, J., Scaramuzza, D.: Learning-based image enhancement for visual odometry in challenging HDR environments. In: 2018 IEEE International Conference on Robotics and Automation (ICRA), pp. 805–811. IEEE (2018)
70. Li, R., Wang, S., Long, Z., Gu, D.: Undeepvo: monocular visual odometry through unsupervised deep learning. In: 2018 IEEE International Conference on Robotics and Automation (ICRA), pp. 7286–7291. IEEE (2018)
71. Gao, X., Zhang, T.: Robust RGB-D simultaneous localization and mapping using planar point features. Robot. Auton. Syst. **72**, 1–14 (2015)
72. Yang, S., Scherer, S.: Direct monocular odometry using points and lines (2017). arXiv preprint arXiv:1703.06380
73. Li, S.J., Ren, B., Liu, Y., Cheng, M.M., Frost, D., Prisacariu, V.A.: Direct line guidance odometry. In: 2018 IEEE International Conference on Robotics and Automation (ICRA), pp. 1–7. IEEE (2018)
74. Wang, T., Ling, H.: Gracker: a graph-based planar object tracker. IEEE Trans. Pattern Anal. Machine Intell. **40**(6), 1494–1501 (2018)

Ming He received his B.Sc., M.Sc. and Ph.D. degrees from PLA Science and Technology University in 2000, 2003 and 2007, respectively. Now he is a professor in the Army Engineering University of PLA. His main research interests focus on emergency command, big data analytics, Internet of Things and public safety.



Chaozheng Zhu received his B.Sc. degree in Computer Science and Technology from Hohai University, Nanjing, China, in 2016. He is currently pursuing his M.Sc. degree in Computer Science and Technology at the Army Engineering University of PLA, Nanjing, China. His research interests include machine vision, visual inertial odometry and embedded systems.

Qian Huang received his B.Sc. degree in Computer Science from Nanjing University, China, in 2003, and Ph.D. degree in Computer Science from the Institute of Computing Technology, Chinese Academy of Sciences, in 2010. From 2010 to 2012, he was a Deputy Technical Manager of Mediatek (Beijing) Incorporation, Beijing, China. Since December 2012, he is with Hohai University, Nanjing, China, where he serves as the Dean of Software Engineering Department. His research interests include multimedia computing, big data analysis, and robot education.

Baosen Ren received his undergraduate degree in Electrical Engineering and Automation from Qingdao University, Qingdao, China, in 2011, and masters degree in Electrical Engineering from Qingdao University in 2015. He is currently working for the State Grid Shandong Electric Power Maintenance Company. The focus of his research is online monitoring and automatic control of power systems.

Jintao Liu received his Ph.D. degree in Control Science and Engineering from the Yantai Aeronautical University, Yantai, China, in 2017. His research interests include machine vision, visual inertial odometry and embedded systems.