Computer Science Review 46 (2022) 100510

Contents lists available at ScienceDirect

Computer Science Review

journal homepage: www.elsevier.com/locate/cosrev

Review Article

Visual SLAM for underwater vehicles: A survey

Song Zhang [a,b], Shili Zhao [a,b], Dong An [a,b], Jincun Liu [a,b], He Wang [a,b], Yu Feng [a,b], Daoliang Li [a,b,c,d,e], Ran Zhao [a,b,*]

[a] *College of Information and Electrical Engineering, China Agricultural University, Beijing 100083, China*
[b] *National Innovation Center for Digital Fishery, China Agricultural University, China*
[c] *Beijing Engineering and Technology Research Centre for Internet of Things in Agriculture, China Agriculture University, Beijing 100083, China*
[d] *China-EU Center for Information and Communication Technologies in Agriculture, China Agriculture University, Beijing 100083, China*
[e] *Key Laboratory of Agricultural Information Acquisition Technology, Ministry of Agriculture, China Agriculture University, Beijing 100083, China*

ARTICLE INFO

Article history:
Received 22 July 2021
Received in revised form 6 April 2022
Accepted 22 August 2022
Available online 11 September 2022

Keywords:
Underwater visual SLAM
Autonomous underwater vehicles
Sensors
Visual odometry
State optimization
Loop closure detection

ABSTRACT

Underwater scene is highly unstructured, full of various noise interferences. Moreover, GPS information is not available in the underwater environment, which thus brings huge challenges to the navigation of autonomous underwater vehicle. As an autonomous navigation technology, Simultaneous Localization and Mapping (SLAM) can deliver reliable localization to vehicles in unknown environment and generate models about their surrounding environment. With the development and utilization of marine and other underwater resources, underwater SLAM has become a hot research topic. By focusing on underwater visual SLAM, this paper reviews the basic theories and research progress regarding underwater visual SLAM modules, such as sensors, visual odometry, state optimization and loop closure detection, discusses the challenges faced by underwater visual SLAM, and shares the prospects of underwater visual SLAM. It is found that the traditional underwater visual SLAM based on filtering methods is gradually developing towards optimization-based methods. Underwater visual SLAM presents a diversified trend, and various new methods have emerged. This paper aims to provide researchers and practitioners with a better understanding of the current status and development trend of underwater visual SLAM, while offering help for collecting underwater vehicles intelligence.

Contents

* Corresponding author at: China Agricultural University, 17 Tsinghua East Road, P.O. Box 121, Beijing 100083, China.
E-mail address: ran.zhao@cau.edu.cn (R. Zhao).

https://doi.org/10.1016/j.cosrev.2022.100510
1574-0137/© 2022 Elsevier Inc. All rights reserved.

Nomenclature

SLAM	Simultaneous Localization and Mapping
AUV	Autonomous Underwater Vehicle
LBL	Long Baseline
SBL	Short Baseline
USBL	Ultra Short Baseline
LiDAR	Light Detection and Ranging
IMU	Inertial Measurement Unit
VO	Visual Odometry
DVL	Doppler Velocity Loggers
ICP	Iterative Closest Point
PnP	Perspective-n-Point
SIFT	Scale Invariant Feature Transform
SURF	Speeded Up Robust Features
ORB	Oriented FAST and Rotated BRIEF
DoG	Difference of Gaussian
FAST	Features from Accelerated Segment Test
BRIEF	Binary Robust Independent Elementary Features
RoI	Region of Interest
CLAHE	Contrast Limited Adaptive Histogram Equalization
EKF	Extended Kalman Filter
KF	Kalman Filter
UKF	Unscented Kalman Filter
PF	Particle Filter
SIR	Sampling Importance Resampling
BA	Bundle Adjustment
EIF	Extended Information Filter
IEKF	Iterated Extended Kalman Filter
ASKF-SLAM	Augmented State Kalman Filter SLAM
UWSim	Underwater Simulator
RANSAC	Random Sample Consistency
CANN	Continuous Attractor Neural Network
SVIn	Sonar Visual Inertial
BoW	Bag-of-words
HALOC	Hash-based Loop Closure

1. Introduction

With the development of robot technology, Autonomous Underwater Vehicle (AUV) has become one of the important means of marine resources exploration and exploitation. Accurate positioning and navigation play a critical role in ensuring that underwater vehicles can move stably and complete the tasks successfully. Due to the rapid attenuation of radio signals, including GPS signals, it is difficult to use GPS for autonomous navigation of underwater vehicles in the water. Long Baseline (LBL), Short Baseline (SBL) and Ultra Short Baseline (USBL) [1] and other methods rely on nearby ships and other carriers to transmit signals to underwater vehicles, making them not suitable for long-distance operation. However, the technology of Simultaneous Localization and Mapping (SLAM) [2] can use sensors to collect different data for analysis and processing. Then, based on the results of data analysis and processing, SLAM can estimate the positions of vehicles, thus making the autonomous localization and navigation of the underwater vehicles possible [3]. As an automatic navigation method, SLAM has made great progress over recent years, with wide application in robot [4] and automatic driving [5]. Different from other methods that rely on external information, SLAM only needs its own sensors to obtain real-time information of the surrounding environment, and can create maps and locate vehicles without any prior input. Thus, vehicles can complete autonomous navigation and positioning in a real sense in any strange environments [3]. Underwater SLAM plays a more and more important role in underwater vehicles navigation.

According to the types of sensors, underwater SLAM can be divided into Light Detection and Ranging (LiDAR) SLAM, sonar SLAM, and visual SLAM. LiDAR and sonar equipment are too expensive, so they are not suitable for civil robots. LiDAR uses laser to analyze the contour and structure of targets. However, because of the existence of small particles, laser will produce absorption and scattering in water, which will affect the measurement results. Therefore, the working range of LiDAR in underwater environment is limited, and maps constructed by LiDAR lack semantic information. Sonar uses a transmitter to emit sound waves and a receiver to receive echo signals, then analyzes and processes the echo signals to describe the contour and structure of the target. Sound wave propagation under water is not affected by light, so sonar is a good choice for underwater SLAM [6]. However, acoustic waves are significantly affected by water flow, seismic activity, ship traffic, marine life and other factors. In addition, in some special cases, such as underwater caves and other closed space, sound waves will be rebounded many times, eventually causing interferences. All of these factors would bring big challenges to underwater positioning and mapping. In contrast, vision-based SLAM has become a hot research field over recent years due to its low cost and high portability, although it can be affected both by particles and light conditions under water. However, various underwater image enhancement algorithms [7] can relieve the difficulties with SLAM to some extent. A comparison of the three SLAM methods is presented in Table 1.

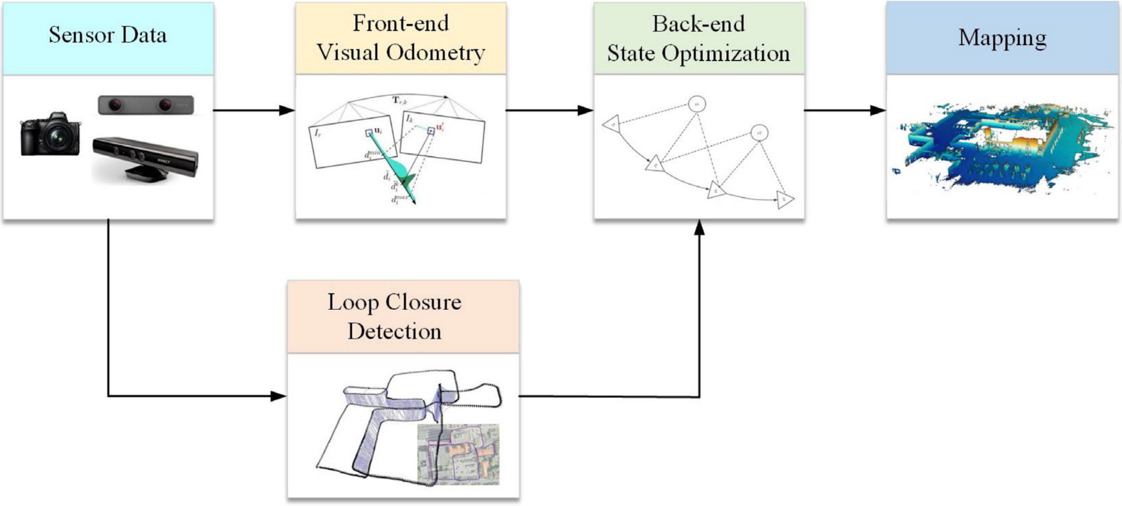


Fig. 1. Framework of visual SLAM.

Table 1
Comparison of the three SLAM methods.

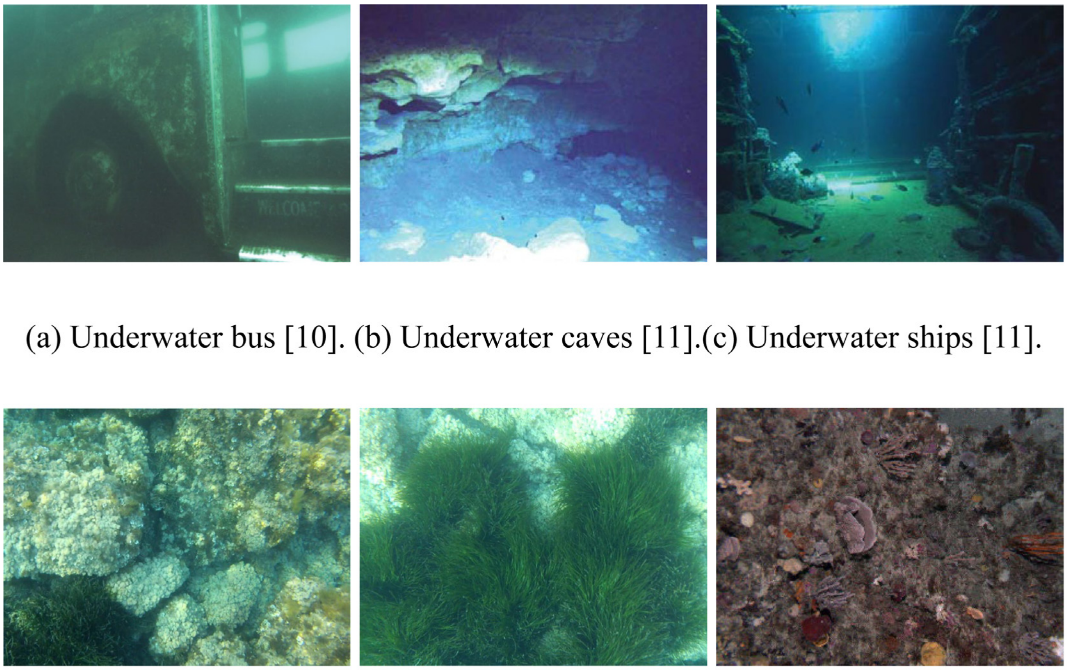
	Advantages	Disadvantages	Applicable scenarios
LiDAR SLAM	High reliability, mature technology, and high precision, with no cumulative error.	With high cost, and limited by radar detection range. The constructed map lacks semantic information.	Mainly used in indoor
Sonar SLAM	It is not affected by light conditions and can work under muddy water.	High cost, and specular reflection that can affect data quality. It is also affected by water flow and seismic wave.	Mainly used underwater
Visual SLAM	Simple structure, easy installation, and low cost. Not limited by the detection distance of sensors.	Affected by environment, it cannot function in dark or textureless areas. It needs heavy computation.	Underwater, indoor and outdoor with good lighting

In order to position and map in unknown environment, sensors shall be used to obtain the key features of the environment, and estimate the current states of vehicles based on the information obtained as well as the previous states of vehicles. As the vehicles keep moving, estimation errors will inevitably appear. To calibrate the errors and ensure the long-term stable work of vehicles, loop closure detection is needed.

The process of visual SLAM can be simply divided into five sections: sensor data, front-end, back-end, loop closure detection, and mapping, as shown in Fig. 1. In visual SLAM, sensors mainly include cameras, as well as some internal sensors of vehicles, such as Inertial Measurement Unit (IMU), depth sensor and so on. The front-end is often called Visual Odometry (VO), which mainly provides optimized sensor data for the back-end system. The back-end optimizes and updates the state of vehicles based on the data from the front-end and loop closure detection, and then calculates the trajectory of vehicles and the map of their surrounding environment. Loop closure detection is used to decide whether a vehicle has reached the previous position and solve the problem of vehicles drift [8] over time.

For static, rigid and unobvious illumination transformation in the scenes without too much interference, the SLAM technology is quite mature [9]. However, different from the ground or indoor controllable environment, underwater environment is highly unstructured, with various kinds of noise interference, which brings multifarious difficulties and challenges to underwater visual SLAM. For instance, due to scattering and absorption, light attenuation exists in water, so the image contrast becomes low. Moreover, the attenuation degrees of different lights in water are different, and it is easier for the lights at higher frequencies to penetrate the particles in water, thus resulting in blue–green underwater images. The dissolved organic matters and suspended particles in water bring huge noise interference. Underwater scenes often show a single structure and lack rich features, which makes it difficult to conduct feature detection and matching. Therefore, underwater SLAM is often much more difficult to implement than on the ground. As shown in Fig. 2, unstructured underwater scenes often include underwater buses, caves, ships, rocks, seaweeds, corals, etc.

Underwater SLAM has attracted wide attention from researchers. For example, [3,10] summarized the state optimization algorithms commonly used in underwater SLAM, and [11] reviewed underwater acoustic SLAM from the perspective of sonar image registration and loop closure detection. Over recent years, underwater visual SLAM has developed rapidly and played an important role in marine resources exploration. However, there is still a lack of systematic review of it. Therefore, after consulting the literature on underwater visual SLAM in recent 15 years on Web of Science, IEEE Xplore and Google Scholar, starting from the framework of visual SLAM, this paper summarizes the development of underwater visual SLAM in recent years, and mainly introduce the following four parts of underwater visual SLAM: related sensors, front-end visual odometry, back-end state optimization, and loop closure detection, as shown in Fig. 3. For positioning, a map can be a simple set of landmarks to meet the requirements of the task. Once the locations of landmarks are determined, the map is constructed. Therefore, this paper will not introduce the mapping process at great length. The structure of the paper is organized as follows. Chapter 1 summarizes the basic situation of underwater SLAM, compares three different underwater SLAM methods, introduces the basic framework of visual SLAM, and highlights the difficulties in special underwater environment; Chapters 2, 3, 4 and 5 introduce the basic content and research status of underwater visual SLAM, covering the related sensors, front-end visual odometry, back-end state



(a) Underwater bus [10]. (b) Underwater caves [11].(c) Underwater ships [11].

(d) Underwater rocks [12]. (e) Underwater seaweeds [12]. (f) Underwater corals [13].

Fig. 2. Underwater unstructured scenes . (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

optimization and loop closure detection, respectively; Chapter 6 discusses difficulties and challenges in the field of underwater visual SLAM; Chapter 7 gives a summary and prospect.

2. Sensors

2.1. Proprioceptive sensors

The sensors for underwater vehicles can be divided into proprioceptive sensors and exteroceptive ones. The latter is mainly used to perceive the external environment, while the former to estimate the state and position of the vehicle itself without external assistance. In addition to necessary external information, the realization of underwater SLAM often requires proprioceptive sensors to provide information such as depth, orientation, and acceleration. Fig. 4 illustrates the common sensors for underwater vehicles, including depth sensor, Doppler Velocity Loggers (DVL), IMU, compass, etc.

DVL works by transmitting sound waves and receiving echoes. According to the Doppler Effect, the velocity of a vehicle can be calculated by comparing the difference of wavelength before and after receiving sound waves. In addition, given that the magnetic poles always point to the north and south of the Earth, the compass provides direction references for vehicles. IMU contains an accelerometer and a gyroscope, which are used to measure acceleration and angular acceleration, respectively. As the vehicles move in water in a three-dimensional manner, the depth information is essential. Depth sensors calculate the depth based on different water pressures at different depths.

2.2. Visual sensors

Visual SLAM mainly uses cameras as the exteroceptive sensor to perceive the external environment information. As a device that forms images following the principle of optical imaging, cameras use photoreceptors to record images. By function, cameras can be divided into mono camera, stereo camera and depth camera.

2.2.1. Mono camera

The monocular adopts only one lens to capture the picture of a target. It has the advantages of simple structure, low cost, and easy calibration and identification. The essence of photo shooting is the projection of a scene on the camera plane, thus getting a two-dimensional expression of the three-dimensional world. As a result, the distance between the target and the camera cannot be calculated from a single picture. Therefore, in order to judge the relative depth of a target, monocular camera often needs to form the disparity through the motion of the camera. This method is also known as the moving view stereo. In addition, the true size of the target in the image cannot be judged by the image alone, and this case is called the scale ambiguity.

2.2.2. Stereo camera

The scale ambiguity and the relative depth can only be calculated based on motion, which poses a greater challenge to SLAM based on mono camera. Stereo camera and depth camera can calculate the depth information more easily. Stereo camera uses more than two cameras, with two of them forming a binocular camera. For a target point in space, its projection coordinates on two camera planes are different. Based on the geometric relationship, the distance between the target and the binocular camera can be calculated according to the coordinates of and the distance between the two cameras (that is, the baseline), and then the spatial position of the target can be obtained. The distance of measurement is affected by the baseline. The longer the baseline, the farther the measurement distance. In essence, the binocular camera uses the parallax between two cameras to achieve ranging, and a lot of calculations are needed to estimate the depth of each pixel. However, the configuration and calibration of binocular cameras are very complex. Other cameras with three or more eyes are essentially similar to binocular cameras.

2.2.3. Depth camera

Depth camera, also known as RGB-D camera, is mainly divided into structured light and time of flight. Its basic mechanism is to

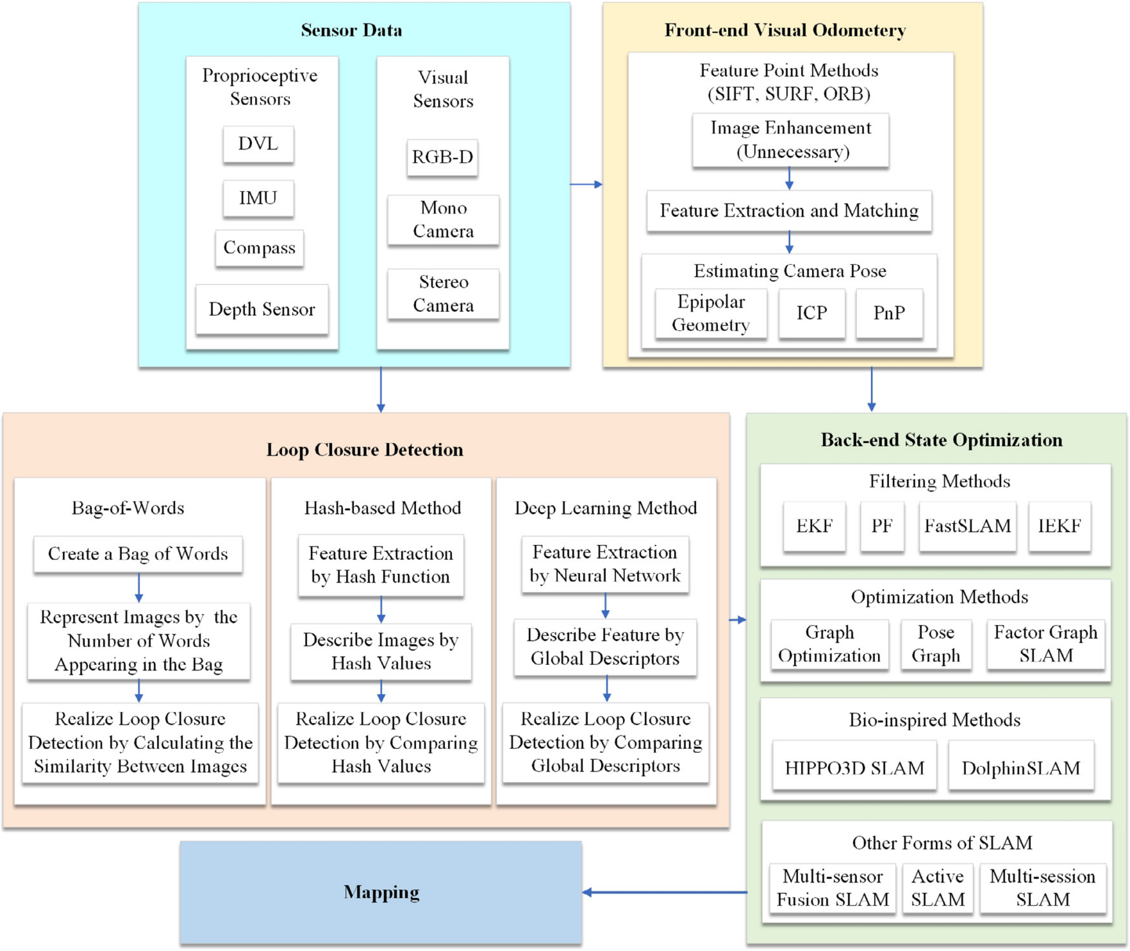

Sensor Data
Proprioceptive Sensors
DVL
IMU
Compass
Depth Sensor
Visual Sensors
RGB-D
Mono Camera
Stereo Camera
Front-end Visual Odometery
Feature Point Methods (SIFT, SURF, ORB)
Image Enhancement (Unnecessary)
Feature Extraction and Matching
Estimating Camera Pose
Epipolar Geometry
ICP
PnP
Loop Closure Detection
Bag-of-Words
Create a Bag of Words
Represent Images by the Number of Words Appearing in the Bag
Realize Loop Closure Detection by Calculating the Similarity Between Images
Hash-based Method
Feature Extraction by Hash Function
Describe Images by Hash Values
Realize Loop Closure Detection by Comparing Hash Values
Deep Learning Method
Feature Extraction by Neural Network
Describe Feature by Global Descriptors
Realize Loop Closure Detection by Comparing Global Descriptors
Back-end State Optimization
Filtering Methods
EKF
PF
FastSLAM
IEKF
Optimization Methods
Graph Optimization
Pose Graph
Factor Graph SLAM
Bio-inspired Methods
HIPPO3D SLAM
DolphinSLAM
Other Forms of SLAM
Multi-sensor Fusion SLAM
Active SLAM
Multi-session SLAM
Mapping

Fig. 3. The framework of underwater visual SLAM.

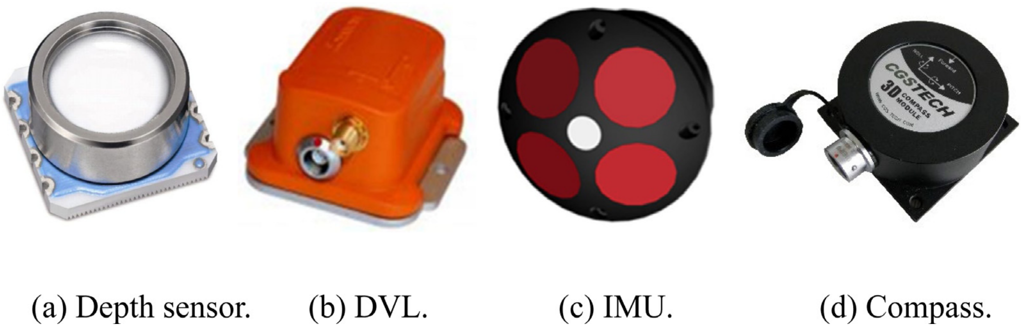


Fig. 4. Proprioceptive sensors for underwater vehicles.

emit a beam of light to the target, and then calculate the distance between the camera and the target by measuring the reflected light. Unlike binocular cameras that use software to calculate distance estimation, depth cameras use physical methods to greatly reduce the computational burden. Due to the features of reflected light, depth cameras have a series of problems such as narrow measurement ranges, high noises, small fields of vision, easy interference by sunlight, and awkward measurement of transmission materials. Implementation of SLAM based on depth camera is generally limited to indoor and other controllable scenes [12,13].

2.3. Summary

Although the commercial RGB-D camera Kinect V1 was released as early as in 2010, the deployment of RGB-D cameras in underwater environments is still very limited [14]. Their main work is around underwater 3D imaging [15] and reconstruction [14], mainly because they adopt infrared light in a large part. However, infrared light will be seriously attenuated in water and has a very limited measurement range. Therefore, it is difficult to use RGB-D cameras as visual sensors to realize underwater visual SLAM. Thus, underwater visual SLAM will still adopt monocular and binocular as visual sensors. The advantages and disadvantages and the representative products of the three kinds of cameras are showcased in Table 2.

3. Front-end visual odometry

The front-end visual odometry is used to roughly estimates the pose of a camera based on the information acquired from

Table 2
Comparison of the three cameras.

	Advantages	Disadvantages	Representative Products
Mono camera	Simple structure and low cost. Can be used indoors or outdoors.	Scale ambiguity. The relative depth is calculated only with parallax formed by motion.	General Camera
Stereo camera	Simple structure and low cost. Can be used indoors or outdoors.	Configuration and calibration are complex, with heavy computation	ZED
RGB-D camera	The depth information is directly measured in physical methods, with accurate measurement	The measurement range is limited and the noise is high, mainly for indoor use.	Kinect/Xtion pro/RealSense

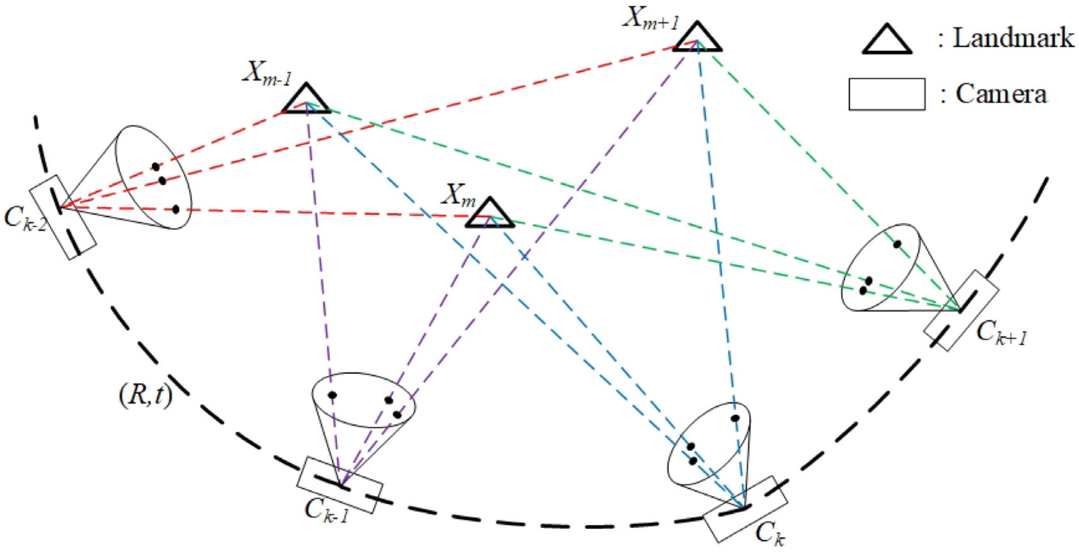


Fig. 5. Schematic diagram of the camera trajectory.

adjacent images, and provides a better initial value for the backend. Visual SLAM without loop closure detection is also called visual odometry. The pose of a camera can be obtained by the following formula:

$$C_k = RC_{k-1} + t \tag{1}$$

where (R, t) is the rotation matrix and translation vector of the camera, at a known initial position C_0, and the camera pose C_k corresponding to any time k can be calculated iteratively. For monocular camera, (R, t) can be solved by epipolar geometry, while for stereo camera and RGB-D camera, (R, t) can be solved by Iterative Closest Point (ICP) [16] and Perspective-n-Point (PnP) [17]. The schematic diagram of the camera trajectory is shown in Fig. 5.

3.1. Feature point method

Due to disturbance and unstable light source, optical flow and direct method in visual SLAM are not available in underwater. For underwater visual SLAM, the front-end calculation method is mainly feature point method. Feature point method matches the representative key points in the images, which has the advantages of stability and insensitivity to illumination and dynamic objects.

Commonly used feature detection and matching methods include Scale Invariant Feature Transform (SIFT) [18], Speeded Up Robust Features (SURF) [19], Oriented FAST and Rotated BRIEF (ORB) [20]. These methods are essentially purposed to find out key points (feature points) on different scales, and have strong robustness to image scaling and rotation transformation. SIFT detects feature points by searching local maxima with Difference of Gaussian (DoG). [21,22] have used SIFT to extract features from images. However, due to its high feature dimension, SIFT consumes heavy computational resources. SURF takes integral images to accelerate feature detection on the basis of SIFT, while retaining the robustness of SIFT. ORB combines the Features from Accelerated Segment Test (FAST) [23] corner detection operator and the Binary Robust Independent Elementary Features (BRIEF) [24] descriptor, and then uses Harris Corner to score the detected corners, so as to filter out the feature points with the highest quality. In the work of Ref. [25], these three feature extraction methods are comprehensively analyzed and summarized, finding that ORB has the fastest computing speed, SIFT and SURF can provide the most scale invariant detectors. Therefore, ORB is recommended for the scenarios with high real-time requirements, and SIFT and SURF are recommended for those with high performance requirements.

To solve the problem of SLAM, it is critical to detect and select the appropriate features, so as to address the data association problem in SLAM [26]. Compared with global images, local images' feature detection and matching can further improve the efficiency. An effective local image feature detection and matching method is to segment the target image into background and Region of Interest (RoI), and then only make detection and matching in the RoI region [27].

3.2. Image enhancement for underwater visual SLAM

Unlike the environment in the air, more and larger particles are suspended in the underwater environment, which leads to the attenuation and scattering of underwater light and the degradation of underwater images. As a result, image feature extraction and matching face difficulties and challenges. To address the problem of image degradation, researchers have proposed many methods to improve the quality of underwater images. Based on different mechanisms, the methods to improve the image quality can be divided into image restoration and image enhancement. Image restoration is used to restore underwater images by simulating the process of underwater image imaging [28]. Image enhancement uses algorithms to directly enhance images and improve visual quality, and certain evaluation indexes are used to supervise this process. Image enhancement is simpler and faster than image restoration [29]. Based on the physical model of imaging, an end-to-end online image enhancement method [30] has been proposed to enhance the visibility of underwater images by introducing artificial light source models. Experiments have shown that the enhanced images can provide more matching feature points and improve the performance of the underwater visual SLAM. In order to reduce the impact of low-quality images on 3D reconstruction and data association, [31] uses homomorphic filtering to normalize images, compensates the uneven illumination area, applies Contrast Limited Adaptive Histogram Equalization (CLAHE) to achieve contrast enhancement, and deploys adaptive denoising filters to reduce the noise generated in the process of contrast enhancement. [32] proposed a hybrid image enhancement method based on model-based and model-free algorithms. For the model-free approach, unsupervised image-processing methods were adopted and modified for fast computation. For the model-based method, particle physics is considered to be applied to synthetic images and real images.

Image enhancement can improve the effect of feature matching to a certain extent. However, given the monotonic structure of underwater scenes and the lack of features, it is still full of challenges to achieve good feature detection and matching.

4. Back-end state optimization

SLAM is essentially an estimation of the uncertainty of the agent itself and the surrounding space [2]. State optimization is the core content of SLAM. Visual odometry gives a short-time pose estimation of cameras, so this process will inevitably lead to cumulative errors. With the accumulation of time, this estimation will become more and more unreliable. On the basis of the visual odometry, the back-end can realize the state optimization in a larger scale and longer time. For underwater visual SLAM, the mainstream state optimization algorithms include the Extended Kalman Filter (EKF) based on filtering theory, the graph optimization and pose graph based on optimization theory, and so on.

4.1. Basic optimization theories and methods

4.1.1. Filtering method

The SLAM problem can be expressed by Bayesian probability distribution function [33]:

$$P(x_k, m|x_0, u_{0:k}, z_{0:k}) \tag{2}$$

where x_k represents the state of a vehicle at k moment; m represents the set of all landmarks; u and z represent input data and observation data, respectively; and subscript $0:k$ indicates all data from 0 to k. Eq. (2) indicates that the joint probability distribution of landmarks and states is related to all input data and observation data from 0 to k. If considering only the relationship between the current time k and the previous time $k-1$, the following formula can be got:

$$P\left(x_k|x_{k-1}, u_k\right) \Longleftrightarrow x_k = f\left(x_{k-1}, u_k\right) + w_k \tag{3}$$

The right side of Eq. (3) is the equation of state, $f()$ represents the motion of the vehicle, and w_k is the random noise. The observation model is as follows:

$$P\left(z_k|x_k, m\right) \Longleftrightarrow z_k = h\left(x_k, m\right) + v_k \tag{4}$$

The right side of Eq. (4) is the observation equation, $h()$ describes the observed geometry, and v_k is the random noise. Eq. (3) indicates that the state x_k at k is only related to x_{k-1} and u_k. Therefore, the state variables can be calculated by iteration and updating. If the state variables obey Gaussian distribution, then only the mean and variance of the state variables need to be maintained. The above two equations describe a basic problem of SLAM: when knowing the motion input u and the observation data (sensor data) z, how to solve the problem of positioning (estimating x) and mapping (estimating m)?

For a linear Gaussian system, its motion equation and observation equation can be expressed as follows:

$$\begin{cases} x_k = A_k x_{k-1} + u_k + w_k \\ z_k = C_k x_k + v_k \end{cases} \qquad k = 1, \ldots, N \tag{5}$$

where A_k and C_k represent the transfer matrix and the observation matrix respectively. It is assumed that all states and noises obey Gaussian distribution, in which the noise obeys zero-mean Gaussian distribution:

$$w_k \sim N\left(0, R\right), \qquad v_k \sim N(0, Q) \tag{6}$$

The function of Kalman Filter (KF) can be divided into two steps: prediction and update. The step of prediction is as follows:

$$\begin{cases} \check{x}_k = A_k \hat{x}_{k-1} + u_k \\ \check{P}_k = A_k \hat{P}_{k-1} A_k^T + R \end{cases} \tag{7}$$

where superscript $\check{x}$ represents the posterior distribution, superscript $\hat{x}$ represents the prior distribution, and P is the covariance matrix of state x. Eq. (7) shows that the state of the previous moment is used to estimate the state of the next moment and the uncertainty (covariance). On the other hand, the step of update is as follows:

$$\begin{cases} \hat{x}_k = \check{x}_k + K(z_k - C_k \check{x}_k) \\ \hat{P}_k = (I - KC_k)\,\check{P}_k \end{cases} \tag{8}$$

where $K = \check{P}_k C_k^T (C_k \check{P}_k C_k^T + Q_k)^{-1}$ denotes the Kalman gain. Eq. (8) denotes that the current state estimation is updated by calculating the weighted average of current measured values.

KF is actually the unbiased estimation of a linear system. However, in the actual SLAM, the system is often nonlinear and cannot be directly used for state optimization. Therefore, linear linearization is often needed. EKF, linearizes the system by first-order Taylor series expansion on the basis of KF, is widely used for underwater visual SLAM [34,35]. Another method is to realize linearization by weighted statistical linear regression, which is called Unscented Kalman Filter (UKF) [36].

Another filtering method based on Bayesian probability is called Particle Filter (PF) [37], which divides the probability in SLAM into a form of factors through Rao–Blackwellized and then estimates the trajectory and map. In the case of trajectory estimation, the Sampling Importance Resampling (SIR) is used to build the map.

4.1.2. Graph SLAM

The filtering method assumes the Markov property to a certain extent, and uses the state at $k-1$ moment to iterate the state at k moment. This kind of "local optimization" will inevitably produce cumulative error. In addition, the first-order Taylor expansion is used for linearization in EKF. When the motion model and observation model have strong nonlinearity, the linearization approximation can only be established in a small range. Furthermore, EKF needs to store and maintain the mean and variance, a process that would consume a lot of memories in large scenes.

Nonlinear optimization methods tend to consider all the historical data, transform the whole SLAM problem into the least squares problem of camera pose and landmarks, and choose the steepest descent method, Newton method and Gauss Newton method to solve the least squares problem. Bundle Adjustment (BA) starts from the observation model (projection process) of the camera, and gets the observation error e as follows:

$$e = z - h(T, p) \tag{9}$$

where $z = h(x, y)$ is the observation equation, T is the Lie group of camera pose x, and P is the three-dimensional point of landmark y. The cost function of BA is obtained as below:

$$\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n}\|e_{ij}\|^2 = \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n}\|z_{ij} - h(T_i, p_j)\|^2 \tag{10}$$

where z_{ij} represents the data generated by the landmark p_j observed at position T_i. The above formula can be solved by the least square method. The independent variable $x = [T_1, T_2, \ldots, T_{m,} p_1, p_2, \ldots, p_n]^T \in \mathbb{R}^{6m}$ is decomposed into pose variable $x_c = [\xi_1, \xi_2, \ldots, \xi_m]^T \in \mathbb{R}^{6m}$ and landmark variable $x_p = [p_1, p_2, \ldots, p_m]^T \in \mathbb{R}^{3n}$. The incremental equation can be solved by the following iteration:

$$\frac{1}{2}\|f(x + \Delta x)\|^2 \approx \frac{1}{2}\|e + F\Delta x_c + E\Delta x_p\|^2 \tag{11}$$

where F and E represent the partial derivatives of camera pose and landmarks, respectively. To acquire the minimum of the above equation, the linear incremental equation should be solved firstly:

$$H\Delta x = g \tag{12}$$

Take Gauss Newton method as an example. H matrix is expressed as follows:

$$H = J^T J = \begin{bmatrix} F^T F & F^T E \\ E^T F & E^T E \end{bmatrix} \tag{13}$$

where $J = [F, E]$ denotes Jacobian matrix.

Given that all the variables are to be optimized, the solution scale of Eq. (12) will become very large. However, H matrix is sparse and can be represented by graph optimization method [38,39]. As shown in Fig. 6, assume that there are two cameras (C_1, C_2) and 5 landmarks $(P_1, P_2, P_3, P_4, P_5)$. Camera C_1 can only observe P_1, P_2, P_3, P_4, while camera C_2 can only observe P_2, P_3, P_4, P_5. This means that the Jacobian matrix J_{ij} corresponding to e_{ij} is only related to pose C_i and landmark P_j, and is 0 in other poses and landmarks. For sparse H matrix, Schur elimination can be used to solve it.

4.1.3. Graph method for pose

In the graph optimization method, landmark nodes and pose nodes will continue to increase with the movement of vehicles, which will bring a huge computational burden. In visual SLAM, landmark nodes are much larger than pose nodes, so the optimization of landmark nodes is abandoned, but only focused on pose node x_c. T represents the pose of the camera, and the relative pose transformation can be calculated according to the pose calculated by the front-end:

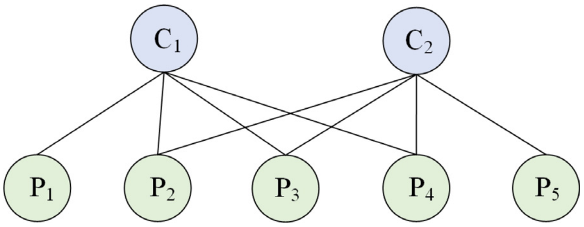


Fig. 6. Diagram of relationship structure.

$$T'_{ij} = T_i^{-1} T_j \tag{14}$$

Based on images i and j, the actual pose increment T_{ij} can be obtained by epipolar geometry. The optimization objectives are as follows:

$$min\, e_{ij} = \min(T_{ij} - T'_{ij}) \tag{15}$$

If the set of all edges is denoted as ε, the overall objective function will be as follows:

$$min \frac{1}{2}\sum_{i,j\in\varepsilon} e_{ij}^T \sum_{ij}^{-1} e_{ij} \tag{16}$$

4.2. Visual SLAM for underwater scenes

4.2.1. Underwater visual SLAM based on filtering methods

In the early underwater visual SLAM, filtering methods were generally used to realize SLAM, including EKF [22,35,40], PF [37], and Extended Information Filter (EIF) [41]. In addition, there are various improved filtering-based methods. For example, based on EKF, the global map is transformed into a local map by using the selective submap joining algorithm [42], thus solving the problem that the amount of calculation would soar with increased maps, and reducing the linear errors [34]. Another way to reduce the linear error is to linearize every EKF iteration, and it is called Iterated Extended Kalman Filter (IEKF) [43]. To improve the accuracy and robustness of underwater vehicles navigation, and reduce the cost of computation, [44,45] have proposed an Augmented State Kalman Filter SLAM (ASKF-SLAM) based on KF. This method stores the vehicle's poses and landmarks in a single state vector, while complementing the prediction and updating of traditional KF by iterating and updating the process state parameters. To realize the vehicle's accurate pose tracking, [46] adopts an improved PF method named FastSLAM2.0 [47], in combination of Bayesian estimation and measurement inversion technology, with vehicle's poses successfully estimated in simulation experiments and real scenes. In the work of [45], FastSLAM2.0 is compared with ASKF-SLAM, finding that ASKF-SLAM have better performance in map management. Table 3 summarizes the underwater visual SLAM based on filtering.

4.2.2. Underwater visual SLAM based on optimization

The optimization algorithm takes into account the historical data of poses and landmarks and delivers good performance in the front of large-scale and long-time scenes [54]. The optimization algorithms mainly include graph optimization and pose graph in underwater visual SLAM [55–57]. In addition, there are also factor graphs, tightly coupled nonlinear optimization, ORB-SLAM [17] and its upgraded version. In Ref. [58], the SLAM problem is modeled with factor graphs, and the external features of vehicles are added to the state as calibration factors; then, the state is optimized by the nonlinear least squares. [59,60] consider

Table 3
Underwater Visual SLAM Based on Filtering.

Reference	Exteroceptive sensors	Environment	Feature detection and matching	State optimization
[22]	Down-looking camera	Stellwagen Bank National Marine Sanctuary	SIFT, and Harris	EKF
[37]	Forward looking camera	Tagiri vent area, and Kagoshima Bay in Japan.	NA	PF
[48]	Stereo camera	Simulated data	SIFT	EKF
[31]	Stereo camera	Simulated data	SIFT, and SURF	EKF
[35]	Mono camera	Experimental water tank, and Stellwagen Bank, National Marine Sanctuary	SIFT	EKF
[41]	Stereo camera	Ningaloo Marine Park	SURF	EIF
[34]	Down-looking camera	Real underwater environment	NA	EKF
[46]	Monocular camera	Simulated data, and sea	SIFT	FastSLAM2.0
[40]	Laser Monocular camera	Portofino, Italy	SURF	EKF
[49]	Monocular camera	Experimental water tank	NA	EKF
[50]	Stereo camera	Simulated data, and real dataset	SURF	EKF
[43]	Stereo camera	Underwater environment simulated by underwater simulator (UWSim) [51], and experimental water tank	SIFT, and Random Sample Consistency (RANSAC)	IEKF
[44]	Monocular camera	Experimental water tank, and sea	SURF	ASKF-SLAM
[52]	Monocular camera	A common area of Port De Valldemossa	SIFT, and RANSAC	IEKF, Multi-session SLAM (refer to 4.2.6)
[53]	Monocular camera	Mallorca coastal area	SITF	IEKF, Multi-session SLAM (refer to 4.2.6)

data from different sensors in the cost function and optimize the entire system through a tightly coupled nonlinear optimization method. Pose constraints were add to the estimation process on the base of ORB-SLAM2 [61]. Based on the ORB-SLAM3, [62] applied the visual-acoustic joint optimization in both tracking and local mapping threads, replacing its original vision only BA.

There are other ways to optimize SLAM in other aspects. From the perspective of entropy, [63] minimizes the entropy of the entire system to maintain the global consistency of the trajectory. Choosing appropriate key frames can effectively diminish the calculation burden and improve the system robustness in the visual SLAM. Starting from image registration, [43] selects key frames by Random Sample Consistency (RANSAC). To address the problem of limited underwater view fields and insufficient features, an online bag of words image saliency measurement method [57] selects key frames according to local saliency and uses global saliency to identify the salient regions in the target. To achieve better 3D view alignment, [31,48] implement Rauch–Tung–Striebel to inversely filter the trajectory calculated by EKF SLAM, so as to obtain a better global consistent trajectory estimation. Table 4 summarizes the underwater visual SLAM based on optimization.

4.2.3. Underwater visual SLAM based on bio-inspired methods

Inspired by biology, researchers proposed a kind of SLAM method from the perspective of bionics, in addition to the above two visual SLAM methods based on filtering and optimization. Considering the limitation of probability method, based on the bionic idea, the paper [66] proposed to use the Continuous Attractor Neural Network (CANN) to simulate the navigation mechanism in mammalian brain, which has successfully extended RatSLAM [67] from 2D ground to 3D underwater environment and developed a HIPPO3D SLAM system. In Ref. [68], a layer of neural network is added to the pose cells in RatSLAM, so as to realize the pose estimation in four degrees of freedom. On the basis of RatSLAM, Ref. [69] leverages neural network to locate and process low resolution visual images and sonar images, and integrates FABMap[1] [70] loop closure detection algorithm into its proposed SLAM system, called DolphinSLAM. In Ref. [71], EKF and DolphinSLAM are compared: On the generated underwater dataset, DolphinSLAM shows similar performance to EKF. Given the shortcomings of EKF in convergence, computational cost and system nonlinearity, DolphinSLAM is supposed to be a better choice. Table 5 summarizes the underwater visual SLAM based on bio-inspired methods.

4.2.4. Underwater SLAM based on multi-sensor fusion

Due to the existence of attenuation and scattering, underwater images often present blurs and color deviations, and the information captured by visual sensors is limited. Therefore, researchers began to use other exteroceptive sensors to integrate visual sensors, so as to deploy underwater SLAM. Sonar images are the visual expression of sonar data and are consistent with optical images in form. Moreover, the processing methods of both are similar. Taking into account the differences between sonar images and optical ones, [69] uses Hessian to detect relatively rich features in optical images, and uses Hu moment to calculate the relatively simple features in sonar images. These features are subsequently fused to realize SLAM of underwater sonar and vision. In Ref. [60], the information on the distance from targets is obtained with mechanical scanning sonars; moreover, the reprojection errors, IMU errors and sonar distance errors are considered in the cost function, and the Sonar Visual Inertial (SVIn) SLAM system is proposed. On this basis, the depth information acquired from pressure sensors is added, the existing SLAM system is optimized, and a new SVIn2 SLAM system is proposed [59]. As a result, the accuracy and robustness of the SVIn2 SLAM system have reached the most advanced level in the test of different underwater visual SLAM datasets. [61] improved the robustness of camera pose estimation in underwater environments by leveraging acoustic odometry. [62] proposed a novel visual-acoustic joint optimization which leverages motion estimates from both DVL and vision to formulate a visual-acoustic residual apart from the reprojection errors. And the calibration of the acoustic odometry was realized by integrating the data from the camera, DVL and depth sensor into the acoustic odometry. In the work of [72], several open source VO packages were compared. The results confirm that incorporating IMU measurements

[1] A probabilistic framework based on Bag-of-Words.

Table 4
Underwater visual SLAM based on optimization.

Reference	Exteroceptive sensors	Environment	Feature detection and matching	State optimization
[63]	Stereo camera	Underwater sunken ship	Direct method	Entropy minimization
[21]	Mono camera	Around the ship at sea	SIFT, and Harris	Pose graph
[57]	Mono camera	Real dataset	SIFT, and SURF	Pose graph
[64]	Monocular camera	Simulated data, a narrow basin on the seafloor, around a ship at sea	NA	Pose graph Active SLAM (refer to 4.2.5)
[65]	Monocular camera	Simulated data, around a ship at sea	NA	iSAM[a] Active SLAM (refer to 4.2.5)
[55]	Mono camera	Experimental water tank	SIFT	Graph optimization
[59]	Mechanical scanning sonar, and stereo camera	Underwater bus, underwater cave, and a fake underwater cemetery	NA	Tightly coupled nonlinear optimization
[60]	Stereo camera, and mechanical scanning sonar	Underwater sunken ship, underwater cave, and underwater bus	NA	Tightly coupled nonlinear optimization
[58]	Monocular camera	Experimental water tank	NA	Factor graph
[56]	Monocular camera	Around a ship at sea	SURF	Pose graph
[61]	Stereo camera, sonar	Water tank, the FloWave Ocean Energy Research Facility in UK	ORB	Improved ORB-SLAM2
[62]	Stereo camera, sonar	Open sea, the FloWave Ocean Energy Research Facility in UK	ORB	Improved ORB-SLAM3

[a] An open frame using factor graph.

Table 5
Underwater visual SLAM Based on bio-inspired methods.

Reference	Exteroceptive sensors	Environment	Feature detection and matching	State optimization
[66]	Monocular camera	Simulated underwater environment with characteristics of oil and gas production field	SURF	HIPPO3D SLAM
[68]	Monocular camera	Simulated underwater environment	SURF	CANN
[69]	Monocular camera, and imaging sonar	Underwater environment simulated by UWSim Seaside port	Hessian for optical image Hu moment for sonar image	DolphinSLAM
[71]	Monocular camera	Underwater environment simulated by UWSim	NA	EKF, and DolphinSLAM

drastically lead to higher performance, in comparison to the pure VO packages.

4.2.5. Active SLAM

Autonomous navigation of underwater vehicles involves three basic issues: localization, mapping and path planning. Traditional SLAM methods are usually passive and need to run along a trajectory planned and controlled in advance. Given the influence of trajectories on SLAM, active SLAM takes path planning into account to improve the effect of localization and mapping by vehicle motions [73]. In order to reduce the uncertainty in navigation, [64] combines sampling-based planning with EIF [41], while taking Gaussian process to predict the environmental saliency of the unpainted areas to conduct image registration. Since the trajectory deviation in visual navigation over time may affect the path planning, [65] proposes a solution to the area coverage problem based on active visual SLAM. In this solution, perception-driven navigation is introduced, and a reward mechanism is used to automatically balance exploration and revisit, thus taking into consideration the path planning and SLAM.

4.2.6. Multi-session SLAM

To combine multiple maps robustly in the same coordinate system and realize the long-term stable operation of vehicles, multiple SLAMs can be performed repeatedly in a certain period in the same environment. This method is called multi-session SLAM [74]. Multi-session SLAM uses the overlapping parts between different sessions, and improve the pose and map estimation of an ongoing session with the data of previous sessions [52]. [52,53] have found multi-session loop closure by comparing hash-based global image signatures to generate more robust global trajectories. [75] introduced an opti-acoustic pairwise factor to realize data association and sharing between sonar image and optical image, and completes the constraints of sonar session and camera session. Multi-session SLAM is an effective method to solve the problem of underwater large-scale mapping.

5. Loop closure detection

In the process of vehicle motion, it is inevitable to produce cumulative error (according to Formula (1)) and thus lead to unreliable long-term estimations as well as failures in establishing globally consistent trajectories and maps. Loop closure detection determines whether a vehicle has reached the previous position by calculating the similarity between maps, and transmits the detection information to the back-end for optimization (the diagram of loop closure detection process is shown in Fig. 7). Therefore, loop closure detection is essentially a feedback link using map data comparison algorithm. Unlike the visual odometry, loop closure detection provides the correlation between the current data and historical data, thus eliminating the cumulative error to a great extent. On the one hand, the wrong loop closure participates in the optimization process, which may make the optimization result converge to the wrong result, resulting in the collapse of the whole SLAM system; On the other hand, increasing the correct number of loop closure can improve the optimization effect and reduce the uncertainty of the system. Therefore, it can significantly improve the accuracy and stability of the whole SLAM [76].

5.1. Bag-of-words method

The simplest loop closure detection method is to directly match any two images [77] and then determine the association

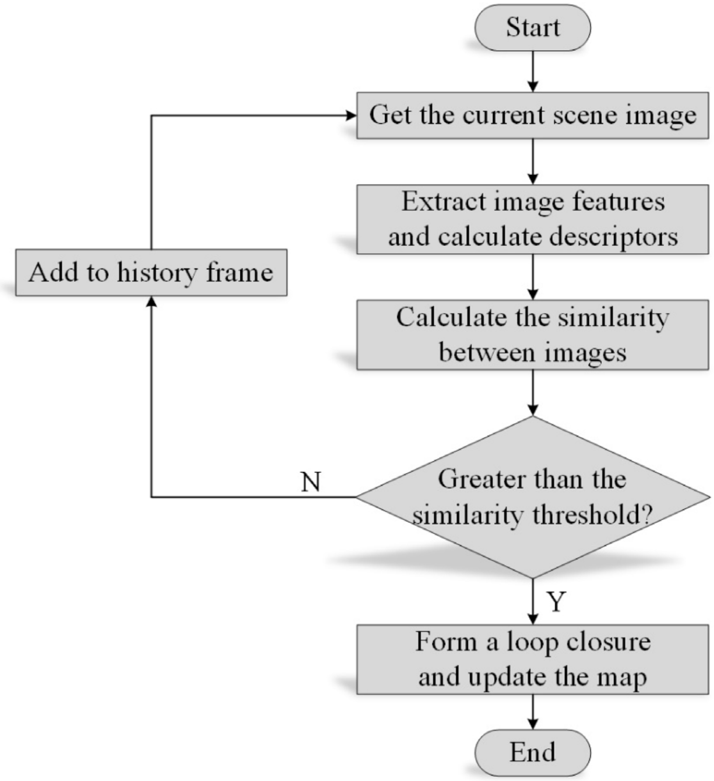

Start
Get the current scene image
Extract image features and calculate descriptors
Calculate the similarity between images
Greater than the similarity threshold?
N
Add to history frame
Y
Form a loop closure and update the map
End

Fig. 7. Diagram of loop closure detection process.

between the images based on the number of matching feature points. This method has some problems: heavy calculation burden, and difficult to meet the real-time requirements of SLAM. An improved method is to randomly extract pictures for comparison [78], but it incurs the problem of low probability as the maps and loops are increased.

The Bag-of-words (BoW) is the most common loop closure detection method. Its basic idea is to use words to represent certain types of features in images, and each image can be represented by a dictionary (a set of words) (the diagram of BoW is shown in Fig. 8). Studies in [44,79] have used the BoW to map the features in images to words in a visual dictionary, and then identify the similarity between the images by counting the words' appearances, so as to decide whether a loop closure case has occurred. By taking advantages of words and dictionaries, only one vector is needed to represent the images, therefore greatly reducing the consumption of computing resources and improving the efficiency and feasibility of loop closure detection.

5.2. Hash-based method

Image hash, also known as media hash, is a global image descriptor used to form fixed length vectors in large data structures. The hash-based method converts the input images into hash values by hash function, and the occurrence of loop closure is judged by comparing the similarity between images' hash values.

Traditional hash algorithms will produce significantly different hash values for small changes in the image, while in visual SLAM, the hash values of similar images are required to be similar [80]. To solve this, Ref. [81] proposes a hash-based loop closure detection method, Hash-based Loop Closure (HALOC), which uses the bucketing mechanism and the unity vector in the projection process to create a unique signature for each group of features, and describes the image through the combination of signatures (as shown in Fig. 9). Since only the calculation of L_1 norm of two vectors is involved, HALOC allows very fast image access. In the comparative experiment with BoW, HALOC has higher recall rate and shorter running time, which greatly reduces the inherent perceptual aliasing problem of BoW and other clustering algorithms [82]. In addition, image hash matching is also used to detect multi-session loop closings, which alleviating the computational cost of the image comparisons [52,53].

5.3. Deep learning method

In the real loops, there are huge amounts of feature categories, but the samples under each feature are very few, thus bringing great challenges to loop closure detection. Many researchers have tried to solve them by utilizing deep learning. Originating from the research of artificial neural network, deep learning is essentially to build a machine learning model with multiple hidden layers to learn valuable features from massive data. For tasks with a large amount of data, machine learning can make outstanding achievements, especially in the fields of image processing [83], natural language processing [84], big data [85] and other domains, due to its strong nonlinear processing ability and feature extracting ability.

In Ref. [86], a loop closure detection method based on the structure of automatic encoder is proposed for underwater visual SLAM. In the encoding part, convolution layer and pooling layer are used to reduce the dimensions of images and extract the main features. In the decoding part, full connection layer converts the feature map to a predefined global image descriptor. Through the comparison of global image descriptors between different images, the loop closure detection of visual SLAM is realized. In Ref. [87], an unsupervised neural network-NetHALOC (as shown in Fig. 10) is proposed based on encoder–decoder structure to convert images into a hash value and then outputs it as a global image descriptor, thus reducing the data operation workload in the process of processing, sharing, comparing and transmitting

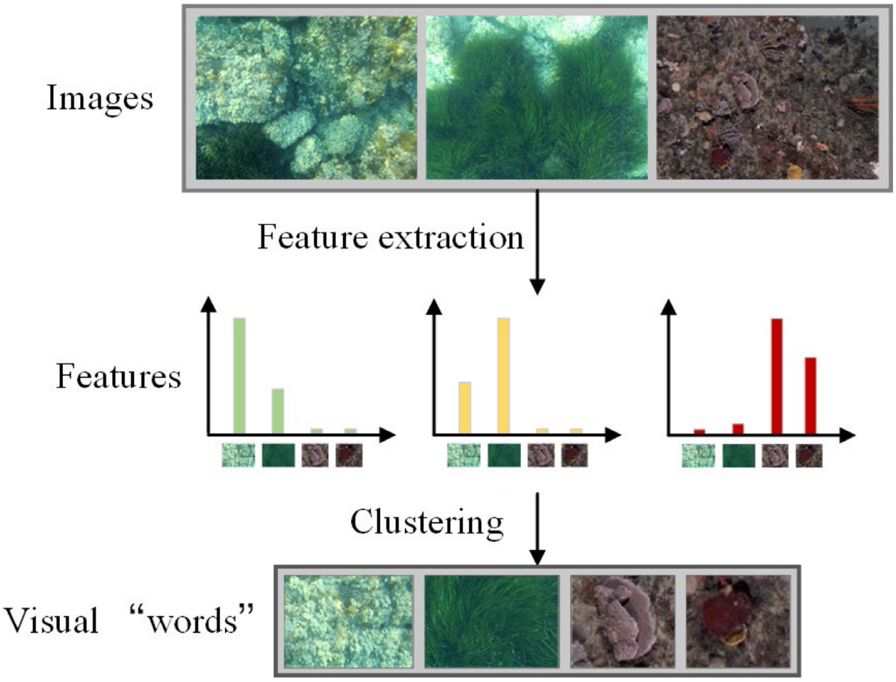


Fig. 8. Diagram of BoW.

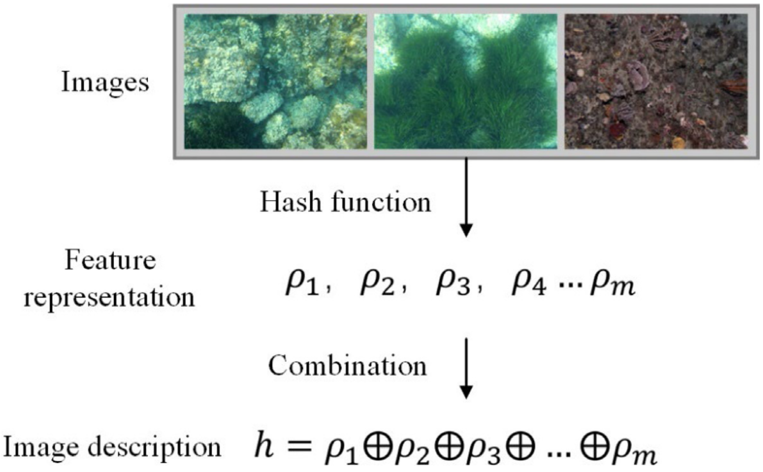

Images
Hash function
Feature representation
$\rho_1, \rho_2, \rho_3, \rho_4 \cdots \rho_m$
Combination
Image description $h = \rho_1 \oplus \rho_2 \oplus \rho_3 \oplus \ldots \oplus \rho_m$

Fig. 9. Diagram of Hash-based method.

images. The method of using hash value to achieve loop closure detection is also called hash matching.

Loop closure detection can be regarded as the problem of recognition and classification: first identifying the features in the current map and then classifying them into categories with similar features. Deep learning shows excellent performance in classification and recognition tasks, and can be used as an effective loop closure detection method.

6. Challenges in underwater visual SLAM

Compared with the laser and sonar methods, visual SLAM is not only cheap, easy to implement and install, but also convenient to create dense maps, because it can capture rich features [88]. For more complex tasks, such as reconstruction [89] and interaction [90], visual SLAM has more advantages. However, underwater is often an unstructured dynamic environment full of various noises, which impose on underwater visual SLAM a lot of great challenges:

1. The sensor data are noisy. As a result, the accuracy of data provided by sensors is low, often accompanied by noise, which would produce a great impact on the final results of SLAM by making the system difficult to converge [68]. On the one hand, the errors of sensors themselves need to be minimized. On the other hand, the interference and noise caused by special underwater environments have to be overcome. Especially, vision sensors are easy to be disturbed by light and turbidity. Therefore, it is an important and difficult task to preprocess and denoise underwater data.

2. Absolute positions are difficult to obtain. SLAM can only obtain the relative positions of underwater vehicles in the environment, because it depends totally on its own sensors to

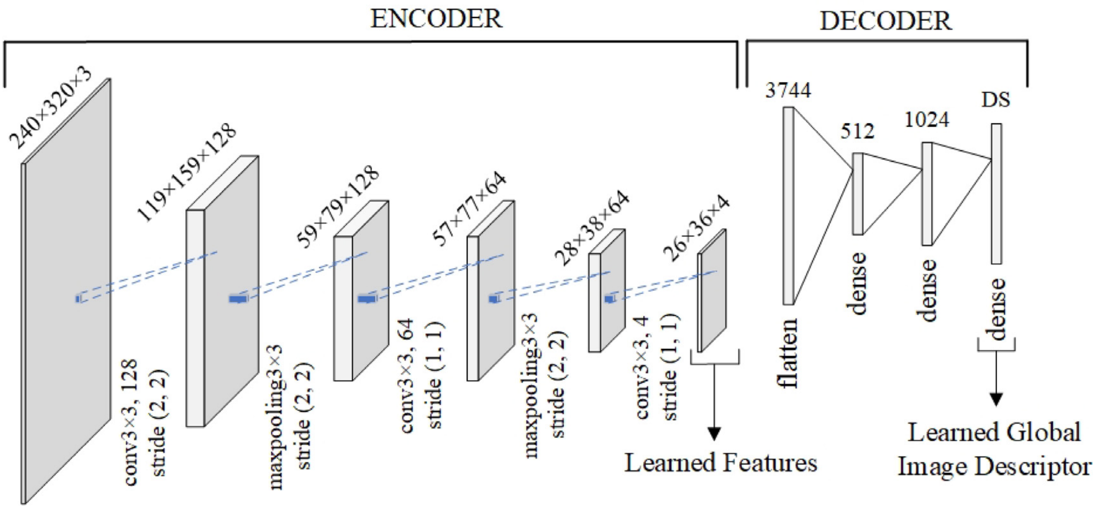


Fig. 10. Diagram of NetHALOC [87].

collect data from the surrounding environment. To obtain absolute positions, it is necessary to draw on other information (GPS, etc.). However, the special underwater environment limits the transmission of GPS signals. Similarly, LBL, USBL, SBL and other alternatives have high installation workload and high cost, while limiting the mobile range of vehicles.

3. Feature extraction is difficult. SLAM needs to effectively identify landmarks from the environment to deliver accurate positioning. Due to the unstructured underwater environment, however, obvious characteristics are lacked in most cases. In addition, underwater light conditions and suspended particles have a great impact on cameras, making the feature extraction more difficult. Even in open waters with good light conditions, floating algae and swimming fish also bring great difficulties and challenges to feature extraction.

4. Robustness and real-time requirements are difficult to meet. Robustness has been a long-term challenge for SLAM [9]. The special underwater environment requires SLAM systems to have certain self-regulation abilities to cope with sudden environmental changes when running for a long time. Furthermore, long-time and large-range movements of vehicles under water will lead to continuously accumulated maps and increased computational complexity, and affect the real-time processing speed of vehicles. On the one hand, the solution to this problem depends on the improvement in the processing capacity of mobile chips; on the other hand, it is necessary to optimize the whole SLAM system to lower the computational complexity.

7. Conclusions and outlook

Starting from the basic framework of visual SLAM, this article explores sensors, front-end visual odometry, back-end state optimization and loop closure detection related to underwater visual SLAM. Further, this article reviews and analyzes the development of underwater visual SLAM in recent years and discusses the existing challenges faced by underwater visual SLAM. Despite this, underwater visual SLAM has made considerable progress and development. Compared with other SLAM solutions, such as sonar and LiDAR, visual SLAM is characterized by low cost and portability, which is the key to reducing the cost of vehicle operations.

1. Although depth cameras can obtain more accurate depth information, most of them are currently equipped with infrared light, which have serious attenuation under water. Therefore, the visual sensors in underwater visual SLAM will still adopt monocular and binocular.

2. Underwater visual SLAM is gradually developing from adopting the traditional filtering-based method to implementing the optimization-based method. Although the optimization-based method provides some obvious advantages in large-scale and long-time scenes [54], the filtering-based method is still effective in the case of limited computing resources on the platform of underwater robots for simple scenes.

3. Light attenuation and suspended particles would degrade the quality of underwater images, thus bringing difficulties to the image feature detection and matching, as well as restricting the deployment of underwater visual SLAM to a certain extent. An effective method is image enhancement, but the gain of this method is limited as well. Underwater environment is complex and diverse. Different environments and tasks require different sensors. Using other sensors together with visual sensors can effectively remedy the defects of visual sensors. In the future, multi-sensor fusion will be an important development direction for underwater SLAM. At the same time, data fusion between different sensors is a great challenge.

4. Looking back on the development of underwater visual LAM in recent years, we find an obvious feature: bio-inspired methods, such as DolphinSLAM and deep learning, are applied in underwater visual SLAM. Not limited to underwater environment, neural network models have been used to detect moving objects, which have been used in semantic visual SLAM [91]. Other similar methods can also be implemented underwater. In the future, more bio-inspired methods will be applied in underwater visual SLAM, so as to promote the intelligent development of underwater SLAM. In addition, in terms of navigation, active SLAM is a more balanced solution, because it takes path planning into account.

5. Data acquisition in the real underwater environment is a complex, expensive and time-consuming task. The publicly available datasets facilitate the testing and validation of the underwater SLAM methods developed so far. With the improvement in information transparency of underwater vehicle industry and the establishment of open databases, researchers will find easier ways to obtain a variety of sample data. At present, there are not many publicly datasets available. Appendix lists some of them that can be downloaded free of charge by researchers.

Table A.1

Reference	Dataset	URL	Description
[92]	AQUALOC	http://www.lirmm.fr/aqualoc/	This dataset is recorded by monocular cameras, Microelectromechanical system IMU, and pressure sensors in three different environments: a harbor at a few meters, a 270 meters-deep archaeological site, and a 380 meters-deep archaeological site.
[51]	Underwater Caves SONAR and Vision Dataset	https://cirs.udg.edu/caves-dataset/	Two mechanical scanning imaging sonars, two IMU and a monocular camera were used to collect data from caves.
[71]	An open set of different simulated datasets using the UWSim	https://goo.gl/GtMQkv	Simulation datasets generated by UWSim. Each dataset contains several tracks in a set of scenes with different turbidity levels.
[93]	Tasmania Coral Point Count	http://marine.acfr.usyd.edu.au/datasets/	A data set of 1258 stereo pairs of the benthos captured by an AUV. Each image has geo-tags from a SLAM solution and 50 expert annotations.
[94]	Scott Reef 25	http://marine.acfr.usyd.edu.au/datasets/	9800 stereo image pairs captured by the Sirius AUV densely covering an area of 75 m × 50 m.
[95]	Tasmania O'Hara 7	http://marine.acfr.usyd.edu.au/datasets/	11,200 stereo image pairs captured by the Sirius AUV traversing a transect of >4 km.
[87]	Mediterranean Dataset	https://github.com/srv/Underwater_Dataset	Two different groups of images were collected, one for the queries (75 images) and the other as a database of images (567 images).

CRediT authorship contribution statement

Song Zhang: Searching and finalizing articles, Formulating research questions, Data extraction, Data cross-checking and analyzing, Writing initial draft, Revising and finalizing article. **Shili Zhao:** Searching and finalizing articles, Revising and finalizing article. **Dong An:** Searching and finalizing articles, Data cross-checking and analyzing, Revising and finalizing article, Supervision. **Jincun Liu:** Searching and finalizing articles, Data cross-checking and analyzing, Revising and finalizing article, Supervision. **He Wang:** Searching and finalizing articles, Revising and finalizing article. **Yu Feng:** Searching and finalizing articles, Revising and finalizing article. **Daoliang Li:** Searching and finalizing articles, Data cross-checking and analyzing, Revising and finalizing article, Supervision. **Ran Zhao:** Searching and finalizing articles, Data cross-checking and analyzing, Revising and finalizing article, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This paper was supported by Ministry of Science and Technology of the People's Republic of China (Grant No. 2019YFE0103700), Hebei Province Department of Science and Technology (Grant No. 20327217D) and Shandong Province Department of Science and Technology (Grant No. 2021TZXD006).

Appendix. Public dataset of underwater visual SLAM

See Table A.1.

References

[1] F. Mandić, I. Rendulić, N. Mišković, Đ. Nađ, Underwater object tracking using sonar and USBL measurements, J. Sens. 2016 (2016).

[2] R.C. Smith, P. Cheeseman, On the representation and estimation of spatial uncertainty, Int. J. Robot. Res. 5 (1986) 56–68.

[3] W. Zhao, T. He, A.Y.M. Sani, T. Yao, Review of SLAM techniques for autonomous underwater vehicles, in: Proceedings of the 2019 International Conference on Robotics, Intelligent Control and Artificial Intelligence, 2019, pp. 384–389.

[4] A. Torres-González, J.R. Martinez-de Dios, A. Ollero, Range-only SLAM for robot-sensor network cooperation, Auton. Robots 42 (2018) 649–663.

[5] K.-W. Chiang, G.-J. Tsai, Y.-H. Li, Y. Li, N. El-Sheimy, Navigation engine design for automated driving using INS/GNSS/3D LiDAR-SLAM and integrity assessment, Remote Sens. 12 (2020) 1564.

[6] A. Palomer, P. Ridao, D. Ribas, Multibeam 3D underwater SLAM with probabilistic registration, Sensors 16 (2016) 560.

[7] M.J. Islam, Y. Xia, J. Sattar, Fast underwater image enhancement for improved visual perception, IEEE Robot. Autom. Lett. 5 (2020) 3227–3234.

[8] M. Labbe, F. Michaud, Online global loop closure detection for large-scale multi-session graph-based SLAM, in: 2014 IEEE/RSJ International Conference on Intelligent Robots and Systems, IEEE, 2014, pp. 2661–2666.

[9] C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. Reid, J.J. Leonard, Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age, IEEE Trans. Robot. 32 (2016) 1309–1332.

[10] F. Hidalgo, T. Bräunl, Review of underwater SLAM techniques, in: 2015 6th International Conference on Automation, Robotics and Applications, ICARA, IEEE, 2015, pp. 306–311.

[11] M. Jiang, S. Song, Y. Li, W. Jin, J. Liu, X. Feng, A survey of underwater acoustic SLAM system, in: International Conference on Intelligent Robotics and Applications, Springer, 2019, pp. 159–170.

[12] Y. Sun, M. Liu, M.Q.-H. Meng, Improving RGB-D SLAM in dynamic environments: A motion removal approach, Robot. Auton. Syst. 89 (2017) 110–122.

[13] Y. Sun, M. Liu, M.Q.-H. Meng, Motion removal for reliable RGB-D SLAM in dynamic environments, Robot. Auton. Syst. 108 (2018) 115–128.

[14] A. Anwer, S.S.A. Ali, A. Khan, F. Mériaudeau, Underwater 3-d scene reconstruction using kinect v2 based on physical models for refraction and time of flight correction, IEEE Access 5 (2017) 15960–15970.

[15] C.-L. Tsui, D. Schipf, K.-R. Lin, J. Leang, F.-J. Hsieh, W.-C. Wang, Using a time of flight method for underwater 3-dimensional depth measurements and point cloud imaging, in: OCEANS 2014-TAIPEI, IEEE, 2014, pp. 1–6.

[16] H. Cho, E.K. Kim, S. Kim, Indoor SLAM application using geometric and ICP matching methods based on line features, Robot. Auton. Syst. 100 (2018) 206–224.

[17] R. Mur-Artal, J.M.M. Montiel, J.D. Tardos, ORB-SLAM: a versatile and accurate monocular SLAM system, IEEE Trans. Robot. 31 (2015) 1147–1163.

[18] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2004) 91–110.

[19] H. Bay, A. Ess, T. Tuytelaars, L. Van Gool, Speeded-up robust features (SURF), Comput. Vis. Image Underst. 110 (2008) 346–359.

[20] E. Rublee, V. Rabaud, K. Konolige, G. Bradski, ORB: An efficient alternative to SIFT or SURF, in: 2011 International Conference on Computer Vision, Ieee, 2011, pp. 2564–2571.

[21] A. Kim, R. Eustice, Pose-graph visual SLAM with geometric model selection for autonomous underwater ship hull inspection, in: 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems, IEEE, 2009, pp. 1559–1565.

[22] R. Eustice, O. Pizarro, H. Singh, Visually augmented navigation in an unstructured environment using a delayed state history, in: IEEE International Conference on Robotics and Automation, 2004. Proceedings. ICRA'04. 2004, IEEE, 2004, pp. 25–32.

[23] E. Rosten, T. Drummond, Machine learning for high-speed corner detection, in: European Conference on Computer Vision, Springer, 2006, pp. 430–443.

[24] M. Calonder, V. Lepetit, C. Strecha, P. Fua, Brief: Binary robust independent elementary features, in: European Conference on Computer Vision, Springer, 2010, pp. 778–792.

[25] S.A.K. Tareen, Z. Saleem, A comparative analysis of sift, surf, kaze, akaze, orb, and brisk, in: 2018 International Conference on Computing, Mathematics and Engineering Technologies (ICoMET), IEEE, 2018, pp. 1–10.

[26] A. Iqbal, N.R. Gans, Data association and localization of classified objects in visual SLAM, J. Intell. Robot. Syst. 100 (2020) 113–130.

[27] J. Aulinas, M. Carreras, X. Llado, J. Salvi, R. Garcia, R. Prados, Y.R. Petillot, Feature extraction for underwater visual SLAM, in: OCEANS 2011 IEEE-Spain, IEEE, 2011, pp. 1–7.

[28] A. Galdran, D. Pardo, A. Picón, A. Alvarez-Gila, Automatic red-channel underwater image restoration, J. Vis. Commun. Image Represent. 26 (2015) 132–145.

[29] R. Schettini, S. Corchs, Underwater image processing: state of the art of restoration and image enhancement methods, EURASIP J. Adv. Signal Process. 2010 (2010) 1–14.

[30] Y. Cho, A. Kim, Visibility enhancement for underwater visual SLAM based on underwater light scattering model, in: 2017 IEEE International Conference on Robotics and Automation, ICRA, IEEE, 2017, pp. 710–717.

[31] J. Salvi, Y. Petillo, S. Thomas, J. Aulinas, Visual slam for underwater vehicles using video velocity log and natural landmarks, in: OCEANS 2008, IEEE, 2008, pp. 1–6.

[32] Y. Cho, A. Kim, Channel invariant online visibility enhancement for visual SLAM in a turbid environment, J. Field Robotics 35 (2018) 1080–1100.

[33] H. Durrant-Whyte, T. Bailey, Simultaneous localization and mapping: part I, IEEE Robot. Autom. Mag. 13 (2006) 99–110.

[34] J. Aulinas, Y.R. Petillot, X. Lladó, J. Salvi, R. Garcia, Vision-based underwater SLAM for the SPARUS AUV, in: Proceedings of the 10th International Conference on Computer and IT Applications in the Maritime Industries. Germany, 2011, pp. 171–179.

[35] R.M. Eustice, O. Pizarro, H. Singh, Visually augmented navigation for autonomous underwater vehicles, IEEE J. Ocean. Eng. 33 (2008) 103–122.

[36] S. Li, P. Ni, Square-root unscented Kalman filter based simultaneous localization and mapping, in: The 2010 IEEE International Conference on Information and Automation, IEEE, 2010, pp. 2384–2388.

[37] T. Maki, H. Kondo, T. Ura, T. Sakamaki, Photo mosaicing of tagiri shallow vent area by the auv tri-dog 1 using a slam based navigation scheme, in: OCEANS 2006, IEEE, 2006, pp. 1–6.

[38] R. Kümmerle, G. Grisetti, H. Strasdat, K. Konolige, W. Burgard, G 2 o: A general framework for graph optimization, in: 2011 IEEE International Conference on Robotics and Automation, IEEE, 2011, pp. 3607–3613.

[39] L. Polok, V. Ila, M. Solony, P. Smrz, P. Zemcik, Incremental block cholesky factorization for nonlinear least squares in robotics, Robot.: Sci. Syst. (2013) 328–336.

[40] F. Ferreira, G. Veruggio, M. Caccia, G. Bruzzone, Real-time optical SLAM-based mosaicking for unmanned underwater vehicles, Intell. Serv. Robot. 5 (2012) 55–71.

[41] I. Mahon, S.B. Williams, O. Pizarro, M. Johnson-Roberson, Efficient view-based SLAM using visual loop closures, IEEE Trans. Robot. 24 (2008) 1002–1014.

[42] J. Aulinas, X. Lladó, J. Salvi, Y.R. Petillot, Selective submap joining for underwater large scale 6-DOF SLAM, in: 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems, IEEE, 2010, pp. 2552–2557.

[43] A. Burguera, F. Bonin-Font, G. Oliver, Towards robust image registration for underwater visual slam, in: 2014 International Conference on Computer Vision Theory and Applications, VISAPP, IEEE, 2014, pp. 539–544.

[44] S. Hong, J. Kim, J. Pyo, S.-C. Yu, A robust loop-closure method for visual SLAM in unstructured seafloor environments, Auton. Robots 40 (2016) 1095–1109.

[45] X. Yuan, J.-F. Martínez-Ortega, J.A.S. Fernández, M. Eckert, AEKF-SLAM: a new algorithm for robotic underwater navigation, Sensors 17 (2017) 1174.

[46] S. Augenstein, S.M. Rock, Improved frame-to-frame pose tracking during vision-only SLAM/SFM with a tumbling target, in: 2011 IEEE International Conference on Robotics and Automation, IEEE, 2011, pp. 3131–3138.

[47] M. Montemerlo, S. Thrun, D. Koller, B. Wegbreit, FastSLAM 2.0: An improved particle filtering algorithm for simultaneous localization and mapping that provably converges, in: IJCAI, 2003, pp. 1151–1156.

[48] J. Salvi, Y. Petillot, E. Batlle, Visual SLAM for 3D large-scale seabed acquisition employing underwater vehicles, in: 2008 IEEE/RSJ International Conference on Intelligent Robots and Systems, IEEE, 2008, pp. 1011–1016.

[49] M. Meireles, R. Lourenço, A. Dias, J.M. Almeida, H. Silva, A. Martins, Real time visual SLAM for underwater robotic inspection, in: 2014 Oceans-St. John's, IEEE, 2014, pp. 1–5.

[50] S. Pi, B. He, S. Zhang, R. Nian, Y. Shen, T. Yan, Stereo visual SLAM system in underwater environment, in: OCEANS 2014-TAIPEI, IEEE, 2014, pp. 1–5.

[51] M. Prats, J. Perez, J.J. Fernandez, P.J. Sanz, An open source tool for simulation and supervision of underwater intervention missions, in: 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, IEEE, 2012, pp. 2577–2582.

[52] A.B. Burguera, F. Bonin-Font, Towards multi session visual SLAM in underwater environments colonized with posidonia oceanica, in: 2018 IEEE/OES Autonomous Underwater Vehicle Workshop, AUV, IEEE, 2018, pp. 1–7.

[53] A. Burguera Burguera, F. Bonin-Font, A trajectory-based approach to multi-session underwater visual slam using global image signatures, J. Mar. Sci. Eng. 7 (2019) 278.

[54] G. Dubbelman, B. Browning, COP-SLAM: Closed-form online pose-chain optimization for visual SLAM, IEEE Trans. Robot. 31 (2015) 1194–1213.

[55] P. Du, J. Han, J. Wang, G. Wang, D. Jing, X. Wang, F. Qu, View-based underwater SLAM using a stereo camera, in: OCEANS 2017-Aberdeen, IEEE, 2017, pp. 1–6.

[56] S. Hong, J. Kim, Three-dimensional visual mapping of underwater ship hull surface using piecewise-planar slam, Int. J. Control Autom. Syst. 18 (2020) 564–574.

[57] A. Kim, R.M. Eustice, Real-time visual SLAM for autonomous underwater hull inspection using visual saliency, IEEE Trans. Robot. 29 (2013) 719–733.

[58] E. Westman, M. Kaess, Underwater AprilTag SLAM and Calibration for High Precision Robot Localization, tech. rep., 2018.

[59] S. Rahman, A.Q. Li, I. Rekleitis, Svin2: an underwater slam system using sonar, visual, inertial, and depth sensor, in: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, IEEE, 2019, pp. 1861–1868.

[60] S. Rahman, A.Q. Li, I. Rekleitis, Sonar visual inertial SLAM of underwater structures, in: 2018 IEEE International Conference on Robotics and Automation, ICRA, IEEE, 2018, pp. 5190–5196.

[61] E. Vargas, R. Scona, J.S. Willners, T. Luczynski, Y. Cao, S. Wang, Y.R. Petillot, Robust underwater visual SLAM fusing acoustic sensing, in: 2021 IEEE International Conference on Robotics and Automation, ICRA, IEEE, 2021, pp. 2140–2146.

[62] S. Xu, T. Luczynski, J.S. Willners, Z. Hong, K. Zhang, Y.R. Petillot, S. Wang, Underwater visual acoustic SLAM with extrinsic calibration, in: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, IEEE, pp. 7647–7652.

[63] J.M. Sáez, A. Hogue, F. Escolano, M. Jenkin, Underwater 3D SLAM through entropy minimization, in: Proceedings 2006 IEEE International Conference on Robotics and Automation, 2006. ICRA 2006, IEEE, 2006, pp. 3562–3567.

[64] S.M. Chaves, A. Kim, E. Galceran, R.M. Eustice, Opportunistic sampling-based active visual SLAM for underwater inspection, Auton. Robots 40 (2016) 1245–1265.

[65] A. Kim, R.M. Eustice, Active visual SLAM for robotic area coverage: Theory and experiment, Int. J. Robot. Res. 34 (2015) 457–475.

[66] L. Silveira, F. Guth, P. Drews, S. Botelho, 3D robotic mapping: A biologic approach, in: 2013 16th International Conference on Advanced Robotics, ICAR, IEEE, 2013, pp. 1–6.

[67] M.J. Milford, G.F. Wyeth, Mapping a suburb with a single camera using a biologically inspired SLAM system, IEEE Trans. Robot. 24 (2008) 1038–1053.

[68] F. Guth, L. Silveira, S. Botelho, P. Drews, P. Ballester, Underwater SLAM: Challenges, state of the art, algorithms and a new biologically-inspired approach, in: 5th IEEE RAS/EMBS International Conference on Biomedical Robotics and Biomechatronics, IEEE, 2014, pp. 981–986.

[69] L. Silveira, F. Guth, P. Drews-Jr, P. Ballester, M. Machado, F. Codevilla, N. Duarte-Filho, S. Botelho, An open-source bio-inspired solution to underwater SLAM, IFAC-PapersOnLine 48 (2015) 212–217.

[70] M. Cummins, P. Newman, FAB-MAP: Probabilistic localization and mapping in the space of appearance, Int. J. Robot. Res. 27 (2008) 647–665.

[71] A.C. Duarte, G.B. Zaffari, R.T.S. da Rosa, L.M. Longaray, P. Drews, S.S. Botelho, Towards comparison of underwater SLAM methods: An open dataset collection, in: OCEANS 2016 MTS/IEEE Monterey, IEEE, 2016, pp. 1–5.

[72] B. Joshi, S. Rahman, M. Kalaitzakis, B. Cain, J. Johnson, M. Xanthidis, N. Karapetyan, A. Hernandez, A.Q. Li, N. Vitzilaios, Experimental comparison of open source visual-inertial-based state estimation algorithms in the underwater domain, in: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, IEEE, 2019, pp. 7227–7233.

[73] Y. Chen, S. Huang, R. Fitch, Active SLAM for mobile robots with area coverage and obstacle avoidance, IEEE/ASME Trans. Mechatronics 25 (2020) 1182–1192.

[74] J. McDonald, Multi-Session Visual Simultaneous Localisation and Mapping, National University of Ireland Maynooth, 2013.

[75] H. Jang, S. Yoon, A. Kim, Multi-session underwater pose-graph SLAM using inter-session opti-acoustic two-view factor, in: 2021 IEEE International Conference on Robotics and Automation, ICRA, IEEE, 2021, pp. 11668–11674.

[76] X.-s. Han, D.-p. Zou, L.-g. Jiang, A geometric information fused loop closing method for visual SLAM, Inf. Technol. 320 (2018) 143–146.

[77] P. Newman, K. Ho, SLAM-loop closing with visually salient features, in: Proceedings of the 2005 IEEE International Conference on Robotics and Automation, IEEE, 2005, pp. 635–642.

[78] F. Endres, J. Hess, J. Sturm, D. Cremers, W. Burgard, 3-D mapping with an RGB-D camera, IEEE Trans. Robot. 30 (2013) 177–187.

[79] C. Gu, Y. Cong, G. Sun, Environment driven underwater camera-IMU calibration for monocular visual-inertial SLAM, in: 2019 International Conference on Robotics and Automation, ICRA, IEEE, 2019, pp. 2405–2411.

[80] F. Bonin-Font, P.L.N. Carrasco, A.B. Burguera, G.O. Codina, LSH for loop closing detection in underwater visual SLAM, in: Proceedings of the 2014 IEEE Emerging Technology and Factory Automation, ETFA, IEEE, 2014, pp. 1–4.

[81] P.L. Negre Carrasco, F. Bonin-Font, G. Oliver-Codina, Global image signature for visual loop-closure detection, Auton. Robots 40 (2016) 1403–1417.

[82] P.-Y. Lajoie, S. Hu, G. Beltrame, L. Carlone, Modeling perceptual aliasing in slam via discrete–continuous graphical models, IEEE Robot. Autom. Lett. 4 (2019) 1232–1239.

[83] D. Shen, G. Wu, H.-I. Suk, Deep learning in medical image analysis, Annu. Rev. Biomed. Eng. 19 (2017) 221–248.

[84] V. Sorin, Y. Barash, E. Konen, E. Klang, Deep learning for natural language processing in radiology—fundamentals and a systematic review, J. Am. Coll. Radiol. 17 (2020) 639–648.

[85] Q. Zhang, L.T. Yang, Z. Chen, P. Li, A survey on deep learning for big data, Inf. Fusion 42 (2018) 146–157.

[86] A. Burguera, F. Bonin-Font, An unsupervised neural network for loop detection in underwater visual SLAM, J. Intell. Robot. Syst. 100 (2020) 1157–1177.

[87] F. Bonin-Font, A. Burguera Burguera, NetHALOC: A learned global image descriptor for loop closing in underwater visual SLAM, Expert Syst. 38 (2021) e12635.

[88] Z. Yan, M. Ye, L. Ren, Dense visual SLAM with probabilistic surfel map, IEEE Trans. Vis. Comput. Graphics 23 (2017) 2389–2398.

[89] C.-H. Yeh, M.-H. Lin, Robust 3D reconstruction using HDR-based SLAM, IEEE Access 9 (2021) 16568–16581.

[90] K. Koide, J. Miura, M. Yokozuka, S. Oishi, A. Banno, Interactive 3D graph SLAM for map correction, IEEE Robot. Autom. Lett. 6 (2020) 40–47.

[91] X. Long, W. Zhang, B. Zhao, PSPNet-SLAM: A semantic SLAM detect dynamic object by pyramid scene parsing network, IEEE Access 8 (2020) 214685-214695.

[92] M. Ferrera, V. Creuze, J. Moras, P. Trouvé-Peloux, AQUALOC: An underwater dataset for visual–inertial–pressure localization, Int. J. Robot. Res. 38 (2019) 1549–1559.

[93] M. Bewley, B. Douillard, N. Nourani-Vatani, A. Friedman, O. Pizarro, S. Williams, Automated species detection: An experimental approach to kelp detection from sea-floor AUV images, in: Proc Australas Conf Rob Autom, 2012.

[94] D.M. Steinberg, S.B. Williams, O. Pizarro, M.V. Jakuba, Towards autonomous habitat classification using Gaussian mixture models, in: 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems, IEEE, 2010, pp. 4424–4431.

[95] A. Friedman, D. Steinberg, O. Pizarro, S.B. Williams, Active learning using a variational dirichlet process model for pre-clustering and classification of underwater stereo imagery, in: 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems, IEEE, 2011, pp. 1533–1539.