Visual SLAM technology for medical applications

Shanhua Chen

School of Electrical Engineering, Zhejiang University, Hangzhou, China

3210104421@zju.edu.cn

Abstract. Simultaneous Localization and Mapping (SLAM) is a method to determine the location of a mobile robot in an unfamiliar environment with external information through the information obtained by the sensor and constructs a three-dimensional map of the unknown environment. There are two main types of SLAM: laser SLAM and visual SLAM. Visual SLAM technology uses image information as the only external information source, that is, to use the camera for pose estimation and map construction. SLAM technology using vision as an external information source is a very active research area with many excellent works. In recent years, visual SLAM has been widely used in the medical field, bringing many new methods. This thesis aims to absorb the extensive work on visual SLAM and its application in the medical field, and introduce their latest progress. And discuss the challenges and possible future development trends of visual SLAM technology in the medical field.

Keywords: SLAM; medical applications; visual SLAM .

1. Introduction

SLAM is a technology that estimates the position and attitude of the sensor in the environment [1]. Slam technology was first proposed by Smith in 1986, and has been widely applied in many areas, including augmented reality, robotics, and medicine [2]. The sensors used between different slam technologies may also be different. Commonly used sensors include laser distance, inertial, and cameras. Among them, SLAM using only cameras has been widely discussed and researched because of its simple sensor configuration but difficult implementation technology. This SLAM technology inputs only the visual information from the camera, which is also called visual SLAM. Visual SLAM algorithms are widely used in robotics, AR, and medical fields [3]. Since the sensor is only a camera, visual SLAM has unique advantages in these fields. For example, the camera is an existing sensor in a smartphone, and the software can obtain camera data and directly construct a 3D map through visual SLAM technology [4]. Especially, in the medical field, because the camera can be made very small, it can safely enter the patient's body for inspection, and it is less harmful to the human body, so the visual SLAM technology has been widely researched and applied in the medical field.

The first part of this article mainly introduces the vision-based SLAM technology. It outlines the main components and partial principles of the visual SLAM technology, which will be helpful for readers who are new to SLAM algorithms and visual SLAM technology. Because this article provides an overview and friendly to readers who are new to visual SLAM algorithms. In the second part, this paper introduces the application of visual SLAM technology in the medical field, outlines some related works, and classifies and summarizes their work. For readers who want to study the application of visual SLAM in the medical field, this article provides the current work progress and difficulties, which will help them understand the relevant information and carry out further work. This paper's third part discusses the challenges and possible future development trends of applying visual SLAM in the medical field. This part can provide readers with a unique perspective and help to inspire readers to conduct further research and innovation in related fields.

2. Composition of the visual Simultaneous localization and mapping

2.1. Sensor data preprocessing

Sensors in visual SLAM include cameras, inertial measurement units (IMUs), etc. The preprocessing operation mainly includes filtering the image (median filtering, Gaussian filtering, etc.), dedistortion and other operations, as shown in Table 1.

The function of filtering the image is mainly to remove the noise in the image and reduce the interference. The main filtering includes Gaussian filtering and median filtering. Dedistortion processing is required after filtering. Distortion mainly includes pincushion distortion and barrel distortion. The lens of the camera is a convex lens. Due to the refraction and loss of the lens, the straight line in the real environment becomes a curve in the image, resulting in distortion.

Table 1. Comparison between different sensors.

Monocular sensor	Binocular sensor	RGB-D
Low cost	Calculation depth	active detection
Unlimited distance	Unlimited distance	good reconstruction
Dimensional uncertainty	Complex configuration	small measuring range
initialization problem	heavy calculation	Susceptible to sunlight and material interference

2.2. front end

The front-end mainly studies how to quantitatively estimate the motion of the camera between frames based on adjacent frame images. The problem of positioning is solved by connecting the motion trajectories of adjacent frames to form the camera carrier's motion trajectory. Then according to the estimated position of the camera at each moment, the position of the spatial point of each pixel is calculated to obtain the map.

2.3. rear end

Only the motion of adjacent frames is calculated in the front end, which will inevitably cause cumulative drift. This is because there is a certain error when estimating the motion between two images each time. After multiple transfers between adjacent frames, the error will gradually be Accumulated, and the deviation between the estimated trajectory and the actual trajectory will become larger and larger. This problem can be optimized through back-end optimization and loopback detection to reduce the accumulation of errors.

Back-end optimization is optimizing the front-end results to obtain the optimal pose estimation. Among them, the principle of the least square method is commonly used: to determine the equation with the "minimum sum of squared residuals".

Loop closure detection is also a way to reduce accumulated errors. It means re-identifying the area that has been mapped. Loop detection can reduce the uncertainty of robot and landmark estimation. Currently, the bag-of-words model is often used to solve this problem.

2.4. Mapping

Mapping is to stitch together the surrounding environment during the robot's movement to obtain a complete map. Different types of maps can be used in different scenarios. For example, sparsely marked maps can only be used for positioning. At the same time, dense maps can also be used for navigation, obstacle avoidance, and reconstruction, and semantic maps are used for interaction, such as VR and other scenarios.

3. The application of visual SLAM in the medical field

3.1. Applications of Visual SLAM in Endoscopy

Endoscopy is the insertion of a flexible tube with a camera and a light source into the body to observe the internal organs of the human body to help doctors make a diagnosis. As one of the most important and widespread applications of visual SLAM in the medical field, using visual slam technology in endoscopy has significantly benefited medical inspection and surgery. One of the major challenges in endoscopy is the lack of accurate endoscope positioning within the patient's body. Visual SLAM addresses this by creating a real-time 3D map of the patient's internal environment through visual information from the endoscope's camera. Below are some works on the application of visual SLAM in endoscopy.

Oscar G. Grasa et al. proposed a monocular SLAM method in 2009, as shown in Figure 1. They applied it to endoscopes, using the endoscope image sequence as the only input of the algorithm to build a 3D map of the abdominal cavity and the pose of the endoscope in real-time. They adopted the EKF+ID+JCBB monocular SLAM method. By observing images collected by a monocular endoscope in the abdominal cavity at 25 Hz, experiments verify that the method achieves real-time performance at this frame rate. The method also supports AR annotation and 3D distance measurement on the 3D map. However, there are also limitations in this research. The assumptions of the algorithm are: 1) rigid scene, 2) smooth motion of the endoscope, 3) low motion clutter, which are usually not true in practice [5].

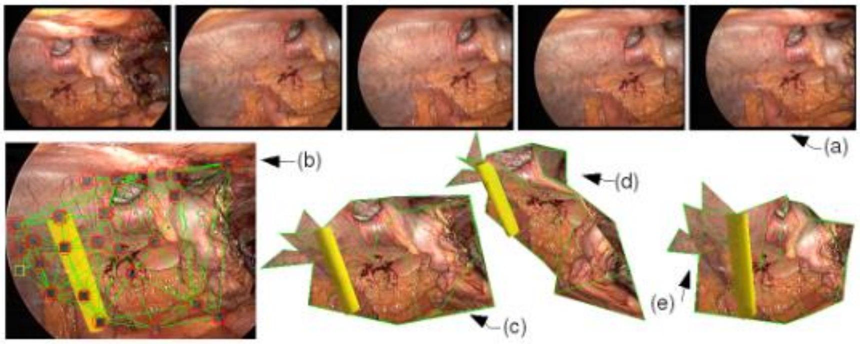

Figure 1. (a) Several frames of endoscopes. (b) Reverse projection of the AR cylinder in the endoscopic video. (c)(d)(e) Insertion of the 3D model and the AR cylinder [5].

Oscar G. Grasa, Javier Civera et al. proposed in 2011 a method combining EKF monocular SLAM+1-point RANSAC+random list repositioning to process laparoscopic image sequences. Based on previous research, the new method proposed by Oscar G. Grasa et al. can be used to deal with more challenging abdominal image sequences. Experiments demonstrate that this approach can address typical challenges in laparoscopic sequences: sudden movements, clutter of surgical tools, temporary tissue deformation, and removal and reinsertion of endoscopes in the abdominal cavity [6].

In 2017, Pawan Kumar Dixit et al. proposed a new method for recovering the 3D shape of polyps from endoscopic video streams through visual SLAM technology. They use sequential forward selection and visual extended Kalman filter to reconstruct real-time 3D model of polyps. This method has been experimentally shown to work well for video sequences without many distorted frames. But the algorithm can't identify distorted or blurred frames, and the algorithm doesn't detect polyps in frames, so when polyps don't appear in some frames in the video sequence, the algorithm gets the wrong data [7].

Xingtong Liu et al. 2022 propose a SLAM system integrating learning-based appearance and optimizeable ensemble priors. The system can reconstruct the dense geometry of anatomical structures from monocular endoscopic video streams. But the current system is only suitable for static scenes [8].

Regine Hartwig et al. 2022 propose a visual slam model to address constrained camera motion. This paper proposes a visual inertial algorithm that can be applied to limited camera motion, such as minimally invasive surgery [9].

Zhuoyue Yang et al. proposed a semantic description method based on scene graph in 2022, as shown in Figure 2. This method combines contour features with SIFT features. Description of this kind of semantic feature can improve the performance of feature matching and increase the precision of pose estimation. Moreover, this research is helpful in developing learning-based approaches. However, this method also has some limitations. If the captured image is blurred or noisy, the matching error will be increased, and the contour extraction may be incomplete [10].

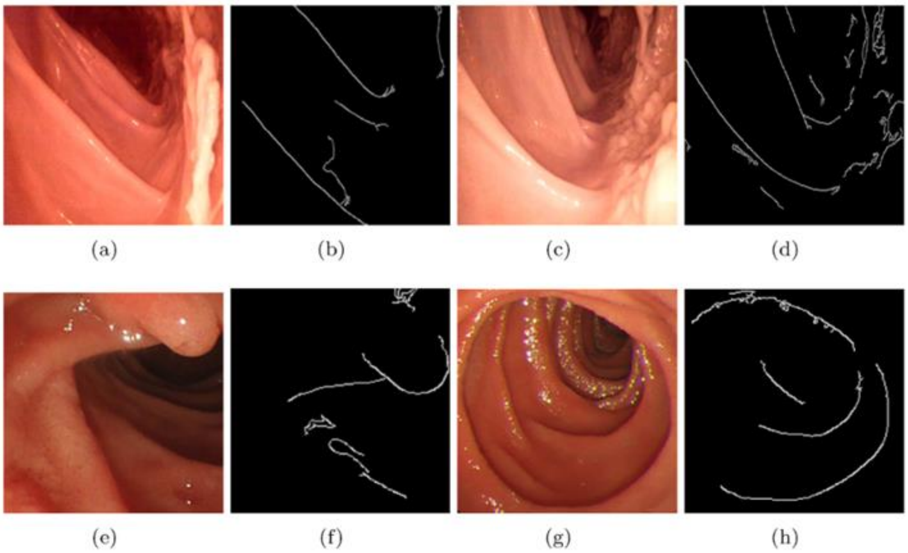

Figure 2. Schematic diagram of contour extraction results in endoscopic images. (a) (c) (e), (g) are the original images, while the rest is an extracted outline. (a) and (c) are from the phantom dataset. (e) and (g) are from the real data set. [10].

3.2. *Application of visual SLAM in intraocular surgery*

Intraocular surgery is performed inside the eye, not on the cornea, and is demanding on doctors because the tiny, delicate structure of the retina requires extreme precision. Vision slam technology can locate retinal blood vessels, helping doctors perform intraocular surgery more precisely. In 2013, Brian C et al. proposed a method for real-time drawing and positioning of retinal blood vessels. This algorithm runs at a frequency of 30Hz in real-time and has a fast convergence speed. However, it may lose track after a large sudden movement, and further adaptation of a more advanced movement model is required [11].

4. Conclusion

This work briefly introduces visual SLAM technology, introduces the main components of visual SLAM and basic principle knowledge. This work also provides an overview of the application of visual SLAM in the medical field. Among them, the application of visual SLAM in the medical field mainly focuses on the application of endoscopy. The application of visual SLAM in endoscopy has the following advantages: 1) Visualize the internal structure of the patient's body during examination and surgery. 2) The 3D map generated by SLAM technology can be combined with AR technology and superimposed on the video source of the endoscope, so that doctors can better understand the spatial relationship inside the target area. 3) With the help of SLAM technology, the endoscope can track its own pose in real time, reducing the dependence on operator skills and the possibility of errors. We summarize and categorize the work of some researchers and list some current challenges and possible directions for future research in visual SLAM in endoscopy. At present, the restricted field of vision, the complex variable structure in the body, and the sudden movement of the endoscope will bring difficulties to the SLAM algorithm. In addition, the real-time performance and computational cost of the system also need to be considered. In the future, researchers can look for more efficient and stable algorithms, so that they can adapt to more complex environments. This work also introduces another visual SLAM applied to intraocular surgery, summarizes the related research work and tries to propose more possible application directions of visual SLAM in the medical field.

References：

[1] A. Tourani, H. Bavle, J. L. Sanchez-Lopez, and H. Voos, “Visual SLAM: What Are the Current Trends and What to Expect?,” Sensors, vol. 22, no. 23, p. 9297, Nov. 2022, doi: 10.3390/s22239297.

[2] A. Macario Barros, M. Michel, Y. Moline, G. Corre, and F. Carrel, “A Comprehensive Survey of Visual SLAM Algorithms,” Robotics, vol. 11, no. 1, p. 24, Feb. 2022, doi: 10.3390/robotics11010024.

[3] T. Taketomi, H. Uchiyama, and S. Ikeda, “Visual SLAM algorithms: a survey from 2010 to 2016,” IPSJ T Comput Vis Appl, vol. 9, no. 1, p. 16, Dec. 2017, doi: 10.1186/s41074-017-0027-2.

[4] J. Fuentes-Pacheco, J. Ruiz-Ascencio, and J. M. Rendón-Mancha, “Visual simultaneous localization and mapping: a survey,” Artif Intell Rev, vol. 43, no. 1, pp. 55–81, Jan. 2015, doi: 10.1007/s10462-012-9365-8.

[5] O. G. Grasa, J. Civera, A. Guemes, V. Munoz, and J. M. M. Montiel, “EKF Monocular SLAM 3D Modeling, Measuring and Augmented Reality from Endoscope Image Sequences”.

[6] O. G. Grasa, J. Civera, and J. M. M. Montiel, “EKF monocular SLAM with relocalization for laparoscopic sequences,” in 2011 IEEE International Conference on Robotics and Automation, Shanghai, China: IEEE, May 2011, pp. 4816–4821. doi: 10.1109/ICRA.2011.5980059.

[7] P. K. Dixit, Y. Iwahori, M. K. Bhuyan, K. Kasugai, and A. Vishwakarma, “Polyp shape estimation from endoscopy video using EKF monocular SLAM with SFS model prior,” in 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), Chennai: IEEE, Mar. 2017, pp. 52–57. doi: 10.1109/WiSPNET.2017.8299718.

[8] X. Liu, Z. Li, M. Ishii, G. D. Hager, R. H. Taylor, and M. Unberath, “SAGE: SLAM with Appearance and Geometry Prior for Endoscopy.” arXiv, Feb. 22, 2022. Accessed: May 15, 2023. [Online]. Available: http://arxiv.org/abs/2202.09487

[9] R. Hartwig, D. Ostler, J.-C. Rosenthal, H. Feußner, D. Wilhelm, and D. Wollherr, “Constrained Visual-Inertial Localization With Application And Benchmark in Laparoscopic Surgery.” arXiv, Feb. 22, 2022.

[10] Z. Yang, J. Pan, R. Li, and H. Qin, “Scene-graph-driven semantic feature matching for monocular digestive endoscopy,” Computers in Biology and Medicine, vol. 146, p. 105616, Jul. 2022, doi: 10.1016/j.compbiomed.2022.105616.

[11] B. C. Becker and C. N. Riviere, “Real-time retinal vessel mapping and localization for intraocular surgery,” in 2013 IEEE International Conference on Robotics and Automation, Karlsruhe, Germany: IEEE, May 2013, pp. 5360–5365. doi: 10.1109/ICRA.2013.6631345.