Proceeding of the 2015 IEEE
International Conference on Information and Automation
Lijiang, China, August 2015

Visual odometry - A Review of Approaches *

Boxin Zhao, Tianjiang Hu and Lincheng Shen
College of Mechatronics and Automation
National University of Defense Technology
Changsha, China, 410073
boxin.zhao@nudt.edu.cn

***Abstract*— As the development of robot autonomous navigation, visual odometry gains more and more attention in the last few years. It can estimate the egomotion of a robot by analyzing the changes of the on-board camera view. This paper gives an overview of visual odometry and the development of its key technologies, such as feature detection, description and matching/tracking, camera pose estimation. The advantages and disadvantages of different methods are compared and the fundamental principles of them are described in this paper.**

***Index Terms*— Visual odometry. Visual SLAM. Robot Localization**

I. Introduction

Visual odometry is a set of methods used to estimate the position and attitude of platforms through analyzing the changes between frame series of camera [14], [17]. For example, I_0 and I_1 are images recorded by camera at time t_0 and t_1, the two images have overlapped areas where contain N pairs of matching feature points $\{(x_i, x_i^{'})|i = 1, 2, ..., N\}$, these corresponding features satisfy the epipolar constraint equation, which is $x_i^{'T} TRx_i = 0$, where T is the antisymmetric matrix of the camera's translation vector, R is the corresponding rotation matrix. Through detecting enough corresponding features, the translation vector T and rotation matrix R of the camera's motion can be obtained. Then, the position of camera at time t_k can be obtained through continuous accumulation of consecutive frames. This is the fundamental principle of conventional visual odometry, involving key technologies such as feature points extraction, description matching or tracking and camera pose estimation. It is not difficult to see from the working principle that since the conventional visual odometry uses information from previous frames and estimates the platform's position of the present moment through continuous accumulation, the positioning accuracy might be affected, namely if there is something wrong with the features matching, the so-caused positioning error will be accumulated to the next time. Therefore, the positioning accuracy cannot be guaranteed in a long time. In order to reduce of this drift, environment mapping through the introduction of 3D information of feature points has become the main solution. The working principle is like this, after matching points between two consecutive frames, positions of the corresponding features in 3D world coordinate system can be calculated based on the triangulation principle. When the next frame is captured, the pose of camera relative to the 3D-point cloud map is calculated, and the map is then updated if new information of feature points are obtained. In this way, information of multiple frames are contained in the map, and the platform-positioning problem has been developed from the pose relationship between two consecutive frames to the pose relationship among multiple previous frames. The key technologies have also been expanded to feature extraction, description matching or tracking, camera pose estimation and map establishment and updating.

II. Feature points detection, description and matching / the research status of tracking technique

A stable feature will not change with the image brightness, rotation, color, size, etc. A great many of researchers have struggled toward this goal for decades, and achieved plentiful and substantial accomplishments. Classic ones, such as the Scale Invariant Feature Transform (SIFT) [6] algorithm, have become a benchmark in the field of computer vision. New algorithms have been constantly proposed afterwards, all of which have set the SIFT algorithm the goal to surpass, such as Speeded Up Robust Features (SURF) [7], [8], Oriented FAST [12] and Rotated BRIEF (ORB) [10] and Binary Robust Invariant Scalable Keypoints (BRISK) [9] . There is also another relatively traditional feature detection algorithm called Harris algorithm [11], the basic idea of which is that if a pixel point is in a flat area, there will be no gray scale variation for the pixel-centered neighboring windows to move towards any direction; if the pixel point is in a edge area, there will be no gray scale variation for its neighboring windows to move towards the edge; if the pixel point is a corner, there will be noticeable gray scale variation for the pixel-centered neighboring windows to move towards any direction. Harris algorithm has eventually introduced a critical equation to determine if a pixel point is an corner or edge point. The feature points extracted by the algorithm have rotation invariance and affine invariance, but no scale invariance. In 2004, in order to explore feature detection operators with scale invariance, Low et al. introduced pyra-

* The authors gratefully acknowledge financial support from Natural Science Foundation of China (No.61403410).

978-1-4673-9104-7/15/$31.00 ©2015 IEEE

mid to construct a multi-scale space, then used Gaussian difference function (DOG operator) to perform detection of extreme points for images of the pyramid's each layer. However, since the calculation of SIFT algorithm has high complexity, it is rarely used in airborne visual odometry that requires high real-time performance. In order to increase the computing speed of SIFT algorithm, discrete Gaussian difference function has been introduced to speed up the detection of feature points. In the 2006 European Conference on Computer Vision (ECCV) conference, a rapid feature points extraction algorithm called FAST was proposed. The idea of FAST algorithm is quite simple, if a point P is a feature point with the gray scale value of I_p, there will be at least N consecutive pixels in its area that have gray scale values greater than (I_p+T) or smaller than (I_p-T), where T is a predefined threshold. FAST algorithm's idea has been used in feature point detection parts of ORB and BRISK algorithms published in the 2011International Conference on Computer Vision (ICCV). Afterwards, in the 2010 ECCV conference, Mair et al. proposed AGAST [3] corner detection algorithm, the idea of which is to apply binary tree into FAST algorithm, and efficiently assign decision trees based on the information of image currently being processed, so as to increase the computing speed of detection operator. Up to now, FAST algorithm and AGAST algorithm have been widely used in visual odmometry because of their exceptional real-time performance and good detection effect.

In order to make sure the detected feature points can find their corresponding points in another image, description matching or tracking is required. Feature point descriptor describes the relationship between a feature point and its neighborhood, thus has certain invariance. For example, rotation invariance is to calculates the direction of a feature point according to the relationship between the feature point and its neighborhood, thus matching points can still be identified when images rotate. In addition, scale invariance and affine invariance are also considered in practical application. All of these are implemented through descriptor. Currently, there are two types of common feature descriptors: multidimensional vector descriptor (such as SIFT, SURF, DAISY, etc.) and binary descriptor (such as BRIEF, BRISK, etc.). SIFT descriptor takes a 16×16 neighborhood around the feature point and divides this neighborhood into 4×4 small areas, 8 directional gradients are recorded for each small area, and eventually a 128-dimension feature vector is used as the descriptor of this point. In recent years, along with researchers' attention to the formation speed of feature point descriptor, binary descriptor has been proposed and widely used. Binary descriptor has abandoned the traditional method that uses gradient histogram to describe area, and randomly picks several pixel point pairs around the feature point, then compares the gray scale values of these points to form a binary string as the feature point's descriptor. The computing speed of binary descriptor is very high, however, since its contained information are considerably lost, it can hardly be applied in image matching of large datasets (such as loop closure), on the contrary, it is more suitable to be used in real-time processing of airborne image sequence.

Be different from the idea of feature point description and matching, feature point tracking detects feature points in one image frame and searches corresponding feature points in the several subsequent frames. This detection-tracking method is usually used for continuous motion problems, the classic method is optical flow algorithm. The motion of objects in real scene will cause the movements of corresponding points on the image, as well as changes in the brightness of corresponding pixels, the brightness change caused by the projection of the motion speed of the pixels' brightness onto the image plane is called optical flow. The basic assumption of feature point tracking based on optical flow method is that the gray scale values of corresponding pixel points in two consecutive image frames are consistent. The bottleneck is how to effectively determine the searching area. In platform motion, in order to effectively estimate the motion's position of attitude, the real-time performance of feature point detection and tracking is highly required. Searching pixel points one by one of the entire image will consume a huge amount of time without being able to guarantee accuracy. Hence, researchers started to introduce motion model of platform [5], for example, airborne units such as IMU and GPS are used to give the possible areas that feature points may appear in the next image frame, thus to increase speed and accuracy of tracking algorithm. Another type of methods use the pyramid theory, in which global searching is performed for low resolution images of the pyramid to determine the approximate positions of feature points, which are then used as the initial values to optimize the positions in high resolution images. Rough features of obvious area are kept in low resolution images, without the interference of detail textures, rapid searching can be implemented. After feature points are located, more detailed information will be introduced in the high resolution images for optimization. This idea was successfully applied in the paper written by Weiss et al. [2] .

In visual odometry, usually image frame sequences are being processed, translational and rotational motions of camera are relatively small, thus it will be counterproductive to use powerful and complex feature point detection and description algorithms, because the process of using descriptor to find matching point actually reduces the information amount of the features point and its neighborhood, therefore, it is not suitable to be used in visual odometry. Comparatively speaking, using image block information with abundant information amount can on the contrary better fit the actual situation. In addition, the constraint of the continuity of camera motion will reduce the searching area of feature point, with certain parameter optimization algorithms to remove outliers, the real-time performance and accuracy of algorithms can be better weighed. The best demonstration data can be found in literatures, in the 2013 Intelligent Robots and Systems (IROS) conference, Weiss et al. used optical flow algorithm to track feature points, and the operating speed on mobile single-core processors was 50 Hz. The following table summarizes classic academic papers in the field of visual odometry over the years, as well as feature point extractions, description matching / tracking algorithms in them. It can be seen from the table data that after years of examination and exploration, even with certain performance defects, FAST, optical flow and other practical algorithms have already been favored by many researchers and applied in all sorts of airborne positioning systems because of their adaptability to environmental noise, low dependence on parameters and high real-time performance.

III. The research status of camera pose estimation technology

Given the key technologies involved in visual odometry and visual Simultaneous Localization and Mapping (SALM), pose relationships between adjacent image frames as well as between 3D map points and camera need to be calculated. The pose relationship between adjacent image frames can be calculated from the projections of one scene on two image frames, and the most common solution is to use epipolar geometric relationship [16]. Epipolar geometric relationship is an analyzing tool for two unmarked images originated from a same scene, this relationship is the only information can be obtained from matching point pairs. Epipolar geometric relationship can be presented using a 3-rank 2-order matrix-basic matrix, it is a data representation of the correspondence between matching point pairs involving all internal and external reference information of the camera. Therefore, through estimation and decomposition of basic matrix, motion parameters of camera (translation vector T and rotation matrix R) can be obtained. The most direct solution is the 8-point algorithm [16] that solves the equation set through finding 8 pairs of matching feature points. This algorithm was proposed by Longuet-Higgins in his master work. The advantage of the 8-point algorithm is that the constructed equation set is linear, thus is faster and requires less computational resources. If there are more than 8 pairs of feature points, they need to be optimized. The commonly used method at present is the RANSAC-8-point algorithm. RANSAC [19] is the abbreviation of random sample consensus algorithm, which mainly uses iterative method to estimate parameters of mathematical models from a set of data containing "outliers". In the field of camera pose estimation, the combination of RANSAC and 8-poing algorithm means selecting 8 pairs of matching feature points through iteration to estimate camera's motion parameters, so that more other feature points satisfy this parameter set. However, although the 8-point algorithm is of simplicity, it is relatively sensitive to noise points. In addition, in order to explore camera pose estimation problems in weak texture environments, researchers have devoted themselves in using fewer points to estimate camera's motion parameters. Many other methods were thus proposed, the classic one is the nonlinear 5-point algorithm [20], which was proposed by Nister in the 2004 Pattern Analysis and Machine Intelligence (PAMI) conference. For this algorithm, given known internal reference of the camera, only 5 pairs of matching points are required to estimate the camera's motion parameters. In 2005, Stewenius analyzed the focus position of camera and proposed the 6-point algorithm [21]. Subsequent relevant literatures separately analyzed related issues of the 5-point algorithm and 6-point algorithm and proposed corresponding improving algorithms. The advantages of these algorithms are that they are implementable linear algorithms, and under most circumstances, output results of these algorithms are very accurate as long as the input feature points are accurate enough, besides, better effects will be achieved if there are more matching points in the image.

In addition to this, more constrains can be introduced to reduce the requirement to the number of matching points. For example, In 2010, Y. Cheng [1] proposed the 4-point algorithm in International Conference on Robotics and Automation (ICRA) conference, in which a coplanar constraint was introduced. In 2011, Weiss et al. introduced Inertial Measurement Unit (IMU) data to assist the calculation of rotation matrix, and subsequently reduced the required number of matching point pairs [2]. In 2011, by using robot model and dynamic constraint, Scaramuzza et al. proposed the 1-point algorithm [18]. Yet, such algorithms have issues such as multiple sulotions and structure degradation, aiming at which, Hartley et al [23]. proposed a nonlinear iterative algorithm in 2009. In this algorithm, the authors used 4 elements and defined a sphere (the radius is) as the rotation parameter space, and the global optimal solution can be search for by traversing the rotation space and branch and bound algorithm. However, the computational burden of this algorithm is quite large. In 2013, Kneip et al. [22] combined the idea of Hartley et al. to optimize the algorithm, and obtained a real-time high-accuracy camera pose estimation algorithm (Eig-based). However, it is no doubt that the advantages of these algorithms are achieved through high computation cost.

The calculation of pose relationship between points in 3D space and camera can be concluded to a series of PnP problems. The problem was first proposed in 1981 and the solution of which has been high developed up to today. The PnP problem is that given coordinates of n points in 3D world coordinate system and their coordinates in camera images, their relative position relationships can be estimated through minimizing objective function $argmin_{T_k} \Sigma_i p_k^i - T_k {X_{k-1}^i}^2$, where X_{k-1}^i is the 3D world coordinate of the ith feature point at the k-1 time point, T_k is the transformation matrix between the world coordinate system and the image coordinate system at k time point, p_k^i is the coordinate of the ith feature point in the image coordinate system at k time point. At present, the application of PnP method is extremely wide, including camera calibration, visual odometry, visual Slam, target tracking, etc. In practical application, n must be $\geq$ 3, then the problem becomes a P3P problem, for which coordinates of at least 3 non-collinear points in the word coordinate system are needed to estimate the camera's relative pose. Rapid solution of P3P problem have been described and studied in many literatures, its C code implementation has also been published, and is the standard method has been mostly applied and studied.

In addition to P3P problem, there are also studies on P4P and P5P problem. The common features shared by these methods are that they are all nonlinear solution and have multiple solutions. Since the used 3D points have very few information, the algorithms are extremely sensitive to errors of image point's position. Hence, in order to increase and robustness of pose estimation results and to improve the anti-noise ability of algorithms, more and more researchers have been considering to introduce a large amount of 3D point information for optimization. A classic example is that Quan et al. considered assigning 3 points into one group, substituting each group into the P3P problem, and eventually solving the results of multiple groups together. The disadvantage of these methods is the greatly increased computational complexity, which in turn reduces the computing speed. In addition, iterative optimization has also been widely applied. For example, the Semi-direct Visual Odometry (SVO) algorithm proposed by Forster et al. in 2014 [5], the Parallel Tracking and Mapping (PTAM) algorithm proposed by Klein et al. [4], all solve the camera pose through iterative optimization. The basic idea is as follows: an objective function is first established so that the 3D point has the minimum error of re-projection in the image, or the gray scale value of the re-projected image point and the gray scale value of the corresponding feature point in adjacent image frame have the minimum difference, the initial values of the camera's pose relative to the world coordinate system are estimated, which can be calculated using relatively simpler methods such as P6P algorithm. The estimated initial values are substituted into the iterative optimization algorithm to constantly optimize the objective function. Such algorithms are the mostly and widely used algorithms in visual odometry and visual Slam. Since they have high computational accuracy and controllable computing speed, with the development of modern optimization theory, these algorithms' performances have been greatly improved. However, since the final optimization results of these methods are closely related to selected iterative optimization algorithm, the algorithm's dependence on initial values, global optimality and convergence all need to be considered in practical application.

IV. Conclusion

This paper described the development history of visual odometry. The traditional visual odometry is the process of estimating the egomotion of a robot by analyzing the changes of the on-board camera. Then in order to reduce the localization drift error, the environment maps are construct. In additional, the key technologies used in visual odometry system are overviewed and compared in this paper.

References

[1] Cheng, Y. (2010), Real-time surface slope estimation by homography alignment for spacecraft safe landing., in 'ICRA' , IEEE, , pp. 2280-2286 .

[2] S. Weiss, D. Scaramuzza, and R. Siegwart, "Monocular-SLAM-Based Navigation for Autonomous Micro Helicopters in GPS-denied Environments," Journal of Field Robotics, vol. 28, 2011.

[3] Elmar Mair, Gregory D. Hager, Darius Burschka, Michael Suppa, and Gerhard Hirzinger. Adaptive and generic corner detection based on the accelerated segment test. In Proceedings of the European Conference on Computer Vision (ECCV'10), September 2010.

[4] Klein G and Murray D. Parallel tracking and mapping for small AR workspaces[C]//Mixed and Augmented Reality, 2007. ISMAR 2007. 6th IEEE and ACM International Symposium on. IEEE, 2007: 225-234.

[5] Forster C and Pizzoli M and Scaramuzza D. SVO: Fast semi-direct monocular visual odometry[C]//Robotics and Automation (ICRA), 2014 IEEE International Conference on. IEEE, 2014: 15-22.

[6] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," in International Journal of Computer Vision (IJCV), vol. 60, no. 2, 2004, pp.91-110.

[7] H. Bay and A. Ess and T. Tuytelaars and L. V. Gool, "SURF: Speeded up robust features," in Computer Vision and Image Understanding (CVIU), vol 110, no. 3, 2008, pp. 346-359.

[8] H. Bay and T. Tuytelaars and L. Van Gool, "SURF: Speeded up robust features," in Proceedings of the European Conference on Computer Vision (ECCV), 2006.

[9] S. Leutenegger and M.Chli and R.Y.Siegwart, "BRISK: Binary Robust Invariant Scalable Keypoints," in Proc. IEEE Int. Conf. Computer Vision (ICCV), 2011.

[10] E. Rublee and V. Rabaud and K.Konolige and G. Bradski, "ORB: an efficient alternative to SIFT or SURF," in Proc. IEEE Int. Conf. Computer Vision (ICCV), 2011.

[11] C. Harris and M. Stephens, "A combined corner and edge detector," in: Alvey Vision Conference. 1988.

[12] E. Rosten and T. Drummond, "Machine learning for highspeed corner detection," in European Conference on Computer Vision (ECCV), volume 1, 2006.

[13] M. Calonder and V. Lepetit and C. Strecha and P. Fua, "BRIEF: Binary Robust Independent Elementary Features," in Proceedings of the European Conference on Computer Vision (ECCV), 2010.

[14] D. Scaramuzza and F. Fraundorfer, "Visual Oodometry: Part I: The First 30 Years and Fundamentals," IEEE Robotics & Automation Magazine, vol 18, no. 4, Dec.2011, pp. 80-92.

[15] A. Alahi and R. Ortiz and P.Vandergheynst, "FREAK: Fast Retina Keypoint," in Proc. Computer Vision and Pattern Recognition (CVPR), 2012.

[16] Y. Ma and S. Soatto and J. Koeck and S.S. Sastry, "An invitation to 3-D vision from images to geometric models," Springer, New York, 2004, pp. 121.

[17] D. Scaramuzza and F. Fraundorfer, "Visual Oodometry: Part II: Matching, Robustness, Optimization, and Applications", IEEE Robotics & Automation Magazine, vol 19, no. 2, Jun. 2012, pp. 78-90.

[18] Davide Scaramuzza, "1-Point-RANSAC Structure from Motion for Vehicle-Mounted Cameras by Exploiting Non-holonomic Constraints", International Journal of Computer Vision, October 2011, Volume 95, Issue 1, pp 74-85.

[19] Martin A. Fischler, Robert C. Bolles, ”Random Sample Consensus: A Paradigm for Model Fitting with Applicationsto Image Analysis and Automated Cartography”, Communications of The ACM - CACM , vol. 24, no. 6, pp. 381-395, 1981.

[20] Nister D., ”An efficient solution to the five-point relative pose problem”, IEEE Trans Pattern Anal Mach Intell. 2004 Jun;26(6):756-77. doi: 10.1109/TPAMI.2004.17.

[21] Stewenius, H. et al., ”A minimal solution for relative pose with unknown focal length”, Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, 2005.

[22] Laurent Kneip, et al. ”direct optimization of frame to frame rotation”, ICCV 2013.

[23] Hartley, R. I.,Kahl, F. (2009). Global Optimization through Rotation Space Search. International Journal of Computer Vision. Springer Netherlands. DOI: 10.1007/s11263-008-0186-9