CCC 2018

Proceedings of the Creative Construction Conference (2018)
Edited by: Miroslaw J. Skibniewski & Miklos Hajdu
DOI 10.3311/CCC2018-136

Creative Construction Conference 2018, CCC 2018, 30 June - 3 July 2018, Ljubljana, Slovenia

Mobile-based 3D Reconstruction of Building Environment

Lei Lei[a,b], Ying Zhou[a,b,*], Hanbin Luo[a,b]

[a]*School of Civil Engineering and Mechanics, Huazhong Univ. of Science and Technology, Wuhan, Hubei 430074, China*
[b]*Hubei Engineering Research Center for Virtual, Safe and Automated Construction, Wuhan, Hubei 430074, China*

Abstract

On-time environment perception of construction site is considered as an indispensable step for project management. Real-time tracking and feedback the status of construction facilitate progress monitoring and quality control. Image-based modelling and RGB-D mapping are considered as a non-invasive and low-cost technology which are always used for data collection and reconstruction of as-built building environments. Recently, the arrival of reliable and efficient computational of mobile terminal service has given us an opportunity to develop a mobile-based spatial data reconstruction system. Considering the capacity of processing and real-time performance on a mobile device, Oriented FAST and Rotated BRIEF (ORB) features are extracted. The ORB features are used for subsequent procedures, including tracking, mapping, relocalization and loop closing. In contrast to image-based off-line modelling, a real time Simultaneous Localization and Mapping (SLAM) algorithm was utilized to estimate the camera trajectory while reconstruction the building environment. Keyframes selection strategy was proposed to reduce the redundant images and generate a robust and trackable sparse point clouds. The keyframes and sparse point clouds are transferred to a computer for generating dense point clouds, grid reconstruction and texture synthesis. Finally, the reconstruction result will be transferred back to mobile and can be displayed directly on a mobile device. As an initial effort, this paper investigated the potential of live reconstruction of indoor building scenes on an android mobile device. Taking the advantages of operable and portable, the system can be used for data acquisition of as-built information by construction workers.

Peer-review under responsibility of the scientific committee of the Creative Construction Conference 2018.

Keywords: 3D Reconstruction, As-built Data Acquisition, Mobile Terminal, Monocular SLAM, Relocalization.

1. Introduction

Three-dimensional information help mitigate the conflict in a construction project where multiple participants are involved [1]. Rather than tedious documents, the visualization of information such as color, texture, shape can be extracted from images/videos that is conducive to communication and decisions-making for stakeholders. Non-intrusive reconstruction of environment technology was widespread used to acquire ‘as-built’ spatial information. The ‘as-built’ information referred to the representation of construction status that can be used for building inspection [2], construction assets tracking [3] and building refurbishment [4].

There are three automated non-contact spatial technologies: (1) Laser scanning technology which collected 3D coordinate of massive points with high accuracy by Terrestrial Laser Scanner. But, the application limited by its high cost, large file sizes and extra specialized operators [5]. (2) Range image-based technology, which using a Red Green Blue and Depth (RGB-D) camera to acquire depth images, and then generate dense point clouds by fusing depth map with multiple viewpoints. While the range camera has the advantage of reconstructing moving objects for equipment and material management, its accuracy is inversely related to the scanning range [6]. Meanwhile, range camera is

Corresponding author: Author email: ying_zhou@hust.edu.cn

sensitive to illumination that result in cannot be accommodated to outdoor environment. (3) Image-based technology which converted 2D continuous photographs into a point cloud model by estimating camera exterior parameters. Structure from Motion (SfM) algorithm was widely used to recover the camera motion and the structure of the scene [7]. However, the SFM triangulation is constrained to the requirement of at least three interrelated images and overlaps between sequential images. As a result, an amount of redundant images are stored that burden computational cost. Moreover, image-based and RGB-D camera based technology are offline data acquiring pattern that is unable to real time predict the performance of modeling. The corollary in this instance is that the invalid data will cause reconstruction process failure.

Technological advancements in field such as robotics, optics and computer vision have been development in real time visual 3D modeling manipulation. Visual Odometry (VO) algorithm [8] was introduced to estimate the trajectory of a camera traversed through its environment. The system, which updates the new camera post with new image adding, combined with Inertial Measurement Unit (IMU) to obtain reconstruction model on a Project Tango Tablet [9]. But the track of features is liable to lose when the occlusion occurs, that obtained 3D information will be discarded.

Visual Simultaneous Localization and Mapping (VSLAM) [10] used vision information to estimate camera motion, and then stored 3D information in the corresponding map. Some references regarded VSLAM as online version of SFM. Scale Invariant Feature Transform (SIFT) and Speeded Up Robust Features (SURF) features [11] were common used for correspondences matching and the structure of the scene recovering. The features are robust to image scaling and rotation, and good invariance to changes in viewpoint and illumination However, SIFT/SURT-SLAM was limited to its large amounts of computational, and unable to operate in real-time without GPU acceleration. The corollary in this instance is that SIFT/SURT feature was not suitable for building sparse map on a mobile device.

In consideration of aforementioned limitations, an ORB-SLAM method is proposed to generate real time sparse point clouds on a hand-hold mobile device. The strategy of selecting and culling keyframes is implemented to avoid unnecessary redundancy. In each keyframe, ORB features are extracted in each keyframe and the matched features are visualized to ensure the validity of triangulation for online modelling. The remainder of this paper is presented in the follow structure. In section 2, a mobile-based algorithms for 3D sparse map generation and optimization is presented. The keyframes and map information are then transmitted to computer to generate 3D grid model in Section 3. Finally, the contribution of the study is highlighted and future work is anticipated.

2. ORB-SLAM operated on mobile device

To achieve the goal of 3D reconstruction building environment, a smartphone with monocular is leveraged to acquire interactive incremental frames. In this instance, tracking and local mapping steps run on the mobile device. Subsequently, the generated sparse point clouds are transformed to personal computer to ensure the computational efficiency. Followed by the dense point clouds, gird reconstruction and texture synthesis steps. The following subsections describe the workflow in detail. Fig. 1 illustrated all the steps refer to the system.

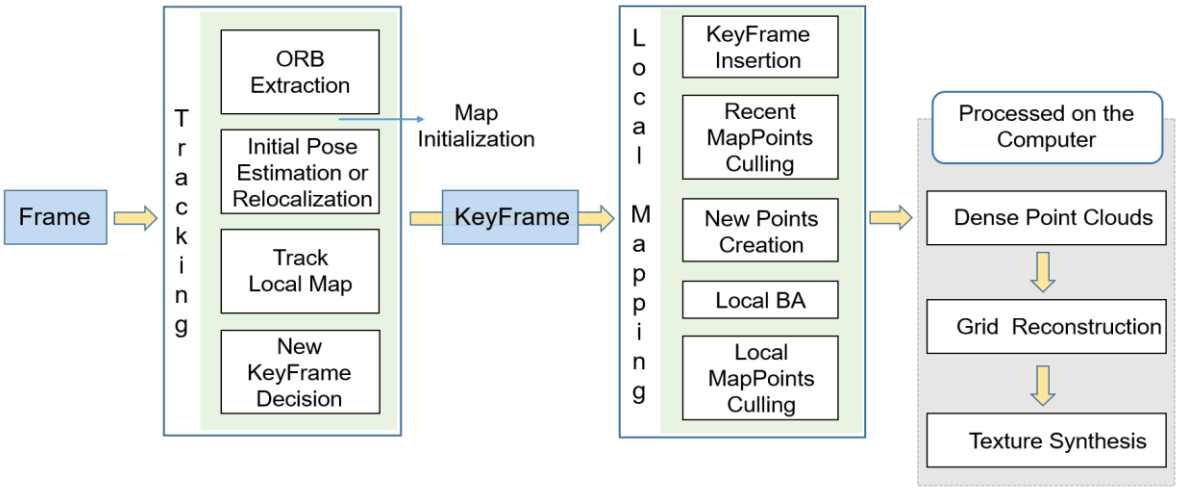


Fig. 1. the workflow for mobile-based reconstruction system.

2.1. Tracking

The process of initialization, tracking and keyframes decision are interpreted in this section. The ORB features are extracted from keyframes for recovering the initial relative camera pose. To ensure consistency of tracking, the bag of words strategy is proposed to speed up correspondences matching in case of tracking lost situation. Meanwhile, the keyframes are selected for the following mapping.

2.1.1. ORB feature extraction

Feature-based Simultaneous Localization and Mapping (SLAM) represents a branch which considering both camera trajectory estimation and environment reconstruction. Oriented FAST and Rotated BRIEF (ORB) feature [12], which is binary feature has good invariance to rotation and scale, built on the FAST key-point detector algorithm [13] and developed BRIEF feature descriptors algorithm [14]. Compared to the SIFT/SURF, ORB features have been greatly improved in extraction speed and efficiency, which allows real-time performance on the mobile device. Moreover, the same features can be used for all subsequent tasks: tracking, local mapping and dense reconstruction.

2.1.2. Initial pose estimation or relocalization

Extracting SFIT or SURF features was expensive in terms of computational time. While BRIEF descriptors take the advantage of fast features matching and comparing which accommodated for operated on the smart phone. The BRIEF descriptor is a binary vector of a square patch around each FAST key point. To find the initial correspondence, extracting features x_1 in the current frame F_1, and searching for matches $x_1 \leftrightarrow x_1'$ in the reference frame F_1'. Computing the Hamming distance between two descriptors of bits and find the correspondence points which with the nearest distance. Considering consistency with previous matches, the camera pose and a guided search of map was predicted based on the last frame. In the case of without enough matching points, a wider search around their position is implemented and the corresponding pose is optimized.

Noteworthy, sudden motions or motion blur is inevitable situation in the process of modelling operations by foremen on the construction site. That would result in map tracking lost and geometric information loss. To overcome the situation, Bag of Words (BoW) [15] is used for matching and place recognition efficiently. A vocabulary tree was built to discretize a binary descriptor space and to speed up geometrical verification. Each image is converted into the bag of words vector and clustered different levels of nodes in the vocabulary tree. The features of images and their associated nodes are stored in the direct index. At the same time, the weight of words in the images are stored in the inverse index. In the context of tracking lost, when adding a new keyframe $I_{f'}$ into the database. The database query to retrieve similar image I_f to the given one. To obtain the correspondences with the highest probability, I_f is searched in the direct index that their features that belongs to the same level nodes. The constraint condition is to speed up the search and matching computation. The RANSAC is performed to calculate a fundamental matrix between I_f and $I_{f'}$, then the camera pose can be found by the PnP (Perspective-n-Point) algorithm [16].

2.1.3. Track local map

After enough matching points are found，8 pair of matching points are selected as a group respectively inside a RANSAC scheme. To reduce computation cost, parallel computing two models: a homography H and a foundation matrix F. The re-projection error and the chi-square distribution are computing a score for each model, record as S_H, S_F. Discarding outliner region and retaining the homograghy and fundamental matrix with the highest score. Calculated the percentage of scores as shown in Eq.(1):

$$R_H = S_H / S_H + S_F \tag{1}$$

If $R_H > 0.45$, which indicates the scene is planar and low parallax, the homography is selected. Otherwise, the fundamental matrix is selected to recover the initial relative camera pose and map.

2.1.4. New keyframe decision

The expansion of scanning range and accumulation of scanning time that resulting in large image file and complexity computation. Considering the extra cost of massive redundant information which indwelled in similar continuous images. Features are extracted from key-frames instead of every frame that reduced the volume of image file and

speeded up data transmission. Obtaining high frame-rates allows to optimize camera tracking and mapping of SLAM system in parallel threads on mobile devices.

To ensure robust relocalization and tracking, and the new keyframe can be inserted as soon as possible. A minimum visual change strategy is imposed. The inserted keyframe should meets the following conditions:

- After the last global Relocalization, more than 20 frames have passed;
- After last keyframe insertion, more than 20 frames have passed;
- At least 50 points in current frame tracks;
- Current frame tracks less than 90% points than feature points in last keyframe.

2.2. Local mapping

As the camera moves with the scanning trajectory, new keyframes are added to the system to grow the map. If a match has been found, a new map point is triangulated and inserted into the map. The strategy of recent map point culling is to ensure the map contain fewer outliners. The point which retained in the map meet the conditions of trackable and not wrongly triangulated. That indicated more than 25 percent of point in the frame are visible and the map point can be observed from at least three keyframes. A covisibility graph [17] is used for real time operating in large scale environment which accommodated to construction site. ORB features are triangulated, and to create new map points from collected keyframes in the covisibility graph.

The local bundle adjustment (BA) is implemented to optimize the surrounding of the features and the camera pose. That is ensure all the keyframes connected to the covisibility graph and all the map points can be seen in their keyframes. Considering the memory space on the smartphone is limited, some keyframes are deleted to avoid unbounded growing. For the keyframes in connected keyframes set, which 90 percent of map points can be seen in more than two keyframes in a fixed scale regarded as the redundant keyframes. The process allows to discard redundant keyframes and maintain a compact reconstruction.

3. Experiment implement and results

An experiment is used to demonstrated the ability to reconstruct building environment with the mobile device acquire ‘as-built’ spatial information. The real time tracking and local mapping steps are performed on a mobile device

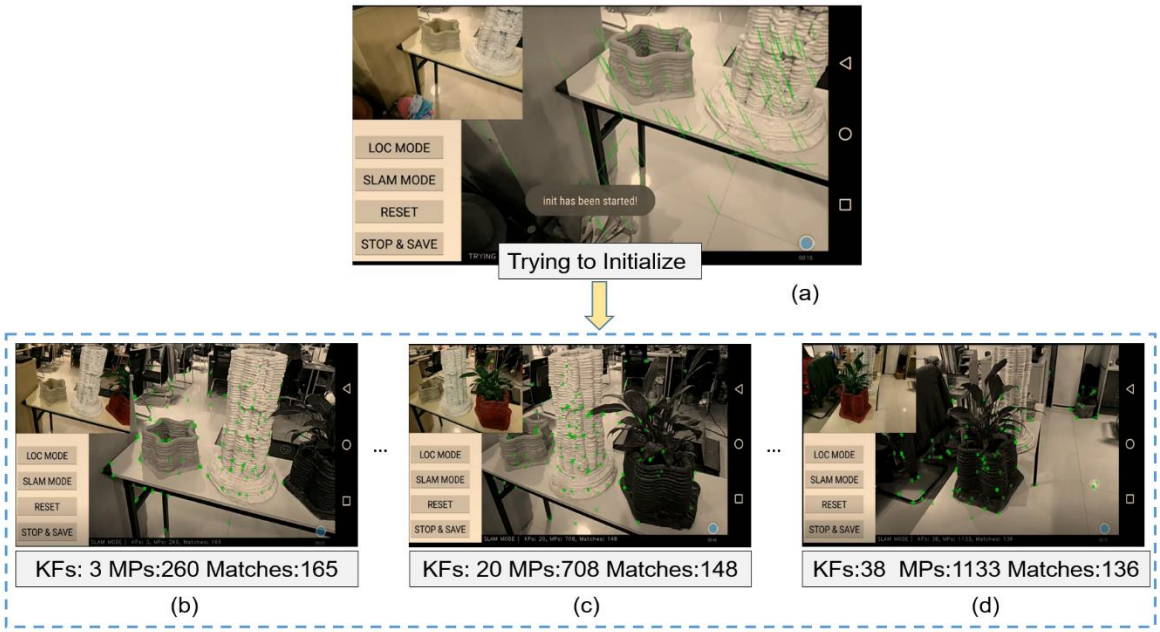


Fig. 2 the real time initialization and keyframes insertion process on the mobile device

with a monocular camera. The result benchmarked on the smartphone with 2.45/1.90 GHz MAM8998 CPU, 6GB RAM, based on Android 7.1 system. The ORB-SALM process of generating keyframes and sparse point clouds data can be seen in Fig. 2. As shown in Fig. 2a, the system trying to estimate the initial camera pose. Once the initialization is completed, the matched points can be found in each keyframe and are denoted in Green Square in Fig. 2b, 2c and 2d. It also shows the current number of keyframe and local map points, respectively represented as 'KFs' and 'MPs'. With the moving of the camera, more keyframes are inserted and new point are triangulated to update the local map points. To ensure the robustness of system and reduce the computational cost, the visualization of sparse point clouds are closed.

A total of 56 keyframes at resolutions of 780*480 pixels and covisibility information are transformed to a computer to enhance computational efficiency. Then, the subsequent steps are performed on a server equipped with a with a 3.10 GHz Intel(R) Xeon(R) E3-1535M CPU, a NVIDIA Quadro P5000 GPU and 64G RAM. Fig. 3 illustrates the process of generating a 3D gird model. A depth propagation algorithm, which is better than conventional PMVS algorithm, is used to transform the sparse point clouds (Fig. 3a) to a dense point clouds, as shown in Fig. 3b. After creating a dense model, *Delaunay* triangulations is used to generate grid triangulations. A graph-cut optimization algorithm is proposed to label each tetrahedron and the gird is extracted from inside and outside the boundary. In view of hallucinations, a small amount of noise is added to eliminate the interference and the result is presented in Fig. 3c. As shown in Fig. 3d, a picture patch of an object is projected back on to the corresponding surface. The pixel of each image patch is assigned a corresponding weight value to improve the clarity for texture synthesis.

For convenience of inspection the reconstruction model on construction site, the 3D grid model is transferred back to the mobile device. Where, Fig. 4 stand for the result of 3D gird reconstruction. In Fig. 4a, the reconstruction result is displayed on a MeshLab platform on the computer. A gird models is displayed via 3D Model Viewer App which is operated on the smartphone as shown in Fig. 4b. It can be seen in Fig. 4b, the performance of gird model on the smartphone can reflect the 3D modelling situation on the computer.

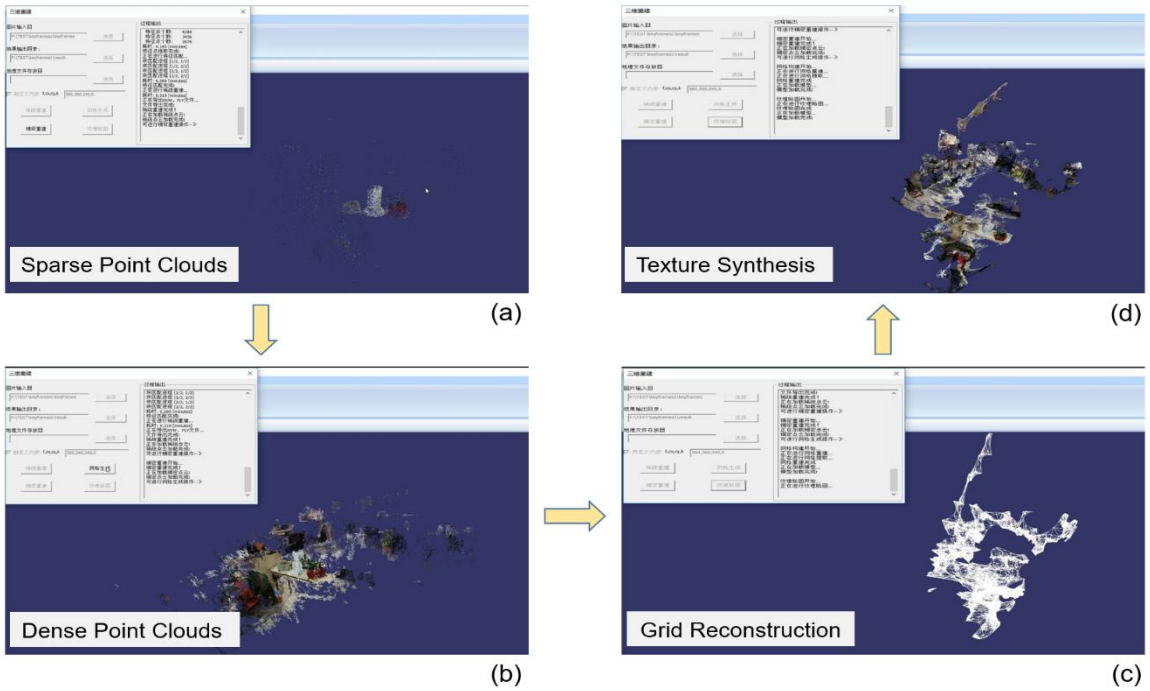
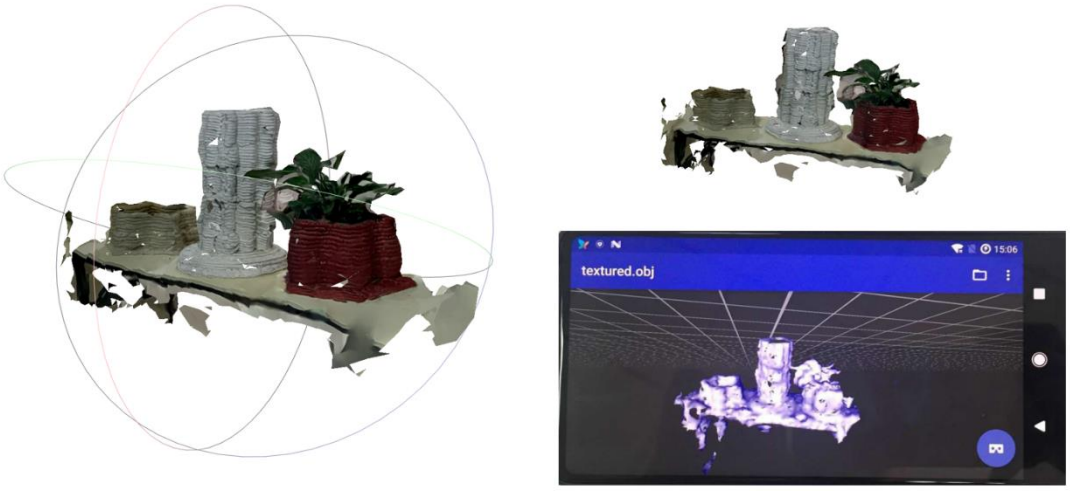


Fig. 3 the process of generating a 3D gird model on a computer



Presented on the Computer (a)

Presented on the Mobile Device (b)

Fig. 4 the results of 3D gird model is presented on a computer screen and a mobile device

4. Conclusion

With the development of smartphone performance, a mobile-based method provide a new way to acquire information on jobsite. To real time visualization the status of building environment，an approach of reconstructing indoor scenes on a hand-hold mobile device is presented in this paper. Capturing 'as-built' spatial data by a smartphone is inexpensive, portable, easy to operate and convenient to inspect for foremen. Oriented FAST and Rotated BRIEF (ORB) feature-based Simultaneous Localization and Mapping (SLAM) is proposed to estimate the camera trajectory and generate a sparse point clouds model on the smartphone. The sparse map transmits to a computer to implement subsequent process for generation a grid 3D model, and then transmits back to the smartphone for inspection.

Future research, however, should focus on real jobsite applications (e.g., large-scale, far-range, poorly textured, etc.) Notably, the comparison will be demonstrated in both laboratory and actual field experiments. The model presented in this paper was reconstructed by using a monocular camera. In order to obtain precise geometry measure of the model, binocular camera will be introduced to complete the reconstruction in future work.

Acknowledgements

This research is supported by National Natural Science Foundation of China (No 71471072).

References

[1] T. Kaneta, S. Furusaka, N. Deng, Overview and problems of BIM implementation in Japan, J. Frontiers of Engineering Management. 2017. 4(2).

[2] D. Lattanzi, G. R. Miller, 3D Scene Reconstruction for Robotic Bridge Inspection, J. Journal of Infrastructure Systems. 2014. 21(2):04014041.

[3] H. Son, C. Kim, 3D structural component recognition and modeling method using color and 3D data for construction progress monitoring, J. Automation in Construction. 2010. 19(7):844-854.

[4] J. Xu, L. Ding, P. E. D. Love, Digital reproduction of historical building ornamental components: From 3D scanning to 3D printing, Automation in Construction 76(2017):85-96.

[5] S. El-Omari, O. Moselhi, Integrating 3D laser scanning and photogrammetry for progress measurement of construction work, J. Automation in Construction. 2009. 18(1):1-9.

[6] Z. Zhu, S. Donia, Spatial and visual data fusion for capturing, retrieval, and modeling of as-built building geometry and features, J. Visualization in Engineering. 2013. 1(1):10.

[7] P. Rodriguez-Gonzalvez, D. Gonzalez-Aguilera, Lopez-Jimenez G, et al, Image-based modeling of built environment from an unmanned aerial system, J. Automation in Construction. 2014. 48:44-52.

[8] J. Engel, J. Sturm, D. Cremers, Semi-dense Visual Odometry for a Monocular Camera, IEEE International Conference on Computer Vision. IEEE. 2014:1449-1456.
[9] T, Schöps, T. Sattler, C. Häne, et al, Large-Scale Outdoor 3D Reconstruction on a Mobile Device, J. Computer Vision & Image Understanding. 2016. 157(C):151-166.
[10] H. Strasdat, J. M. M. Montiel, A. J. Davison, Real-time monocular SLAM: Why filter? IEEE International Conference on Robotics and Automation. IEEE. 2010:2657-2664.
[11] D. G. Lowe, Distinctive Image Features from Scale-Invariant Keypoints, J. International Journal of Computer Vision. 2004. 60(2):91-110.
[12] E. Rublee, V. Rabaud, K. Konolige, et al, ORB: An efficient alternative to SIFT or SURF, J. IEEE Computer Society. 2011. 58(11):2564-2571.
[13] E. Rosten, T. Drummond, Machine learning for high-speed corner detection, European Conference on Computer Vision, Springer-Verlag. 2006:430-443.
[14] M. Calonder, V. Lepetit, C. Strecha, et al, BRIEF: binary robust independent elementary features, European Conference on Computer Vision. 2010:778-792.
[15] D. Galvez-López, J. D. Tardos, Bags of Binary Words for Fast Place Recognition in Image Sequences, J. IEEE Transactions on Robotics. 2012. 28(5):1188-1197.
[16] V. Lepetit, F. Moreno-Noguer, P. Fua, EPnP: An Accurate O(n) Solution to the PnP Problem, J. International Journal of Computer Vision. 2009. 81(2):155-166.
[17] Strasdat, Hauke, et al. Double window optimisation for constant time visual SLAM, International Conference on Computer Vision IEEE Computer Society, 2011:2352-2359.