AN AUTOMATIC KEY-FRAME SELECTION METHOD FOR VISUAL ODOMETRY BASED ON THE IMPROVED PWC-NET

Yu Chen, Li Yan *, Xiaohu Lin

School of Geodesy and Geomatics, Wuhan Univeristy, China
chenyuphd@whu.edu.cn, lyan@sgg.whu.edu.cnf, xhlin214@whu.edu.cn

Commission II, WG II/5

KEY WORDS: VO/VSLAM, key-frame selection, motion state, improved PWC-Net, attitude change.

ABSTRACT:

In order to quick response to the rapid changes of mobile platforms in complex situations such as speedy changing direction or camera shake, visual odometry/visual simultaneous localization and mapping (VO/VSLAM) always needs a high frame rate vision sensor. However, the high frame rate of the sensor will affect the real-time performance of the odometry. Therefore, we need to investigate how to make a balance between the frame rate and the pose quality of the sensor. In this paper, we propose an automatic key-frame method based on the improved PWC-Net for mobile platforms, which can improve the pose tracking quality of odometry, the error caused by dynamic blur and the global robustness. First, a two-step decomposition is used to calculate the change of inter-frame attitude, and then, key-frames are added by the improved PWC-Net or automatically selected based on the motion state of the vehicle predicted by pose change with a short time interval. To evaluate the method, we conduct extensive experiments on KITTI dataset based on monocular visual odometry. The results indicate that our method can keep the pose tracking quality while ensuring the real-time performance.

1. INTRODUCTION

With the development of unmanned aerial vehicle (UAV) technology, autonomous flying, high precision localization and mobile mapping in an unknown environment are important. Recently, VO/VSLAM, as an effective complement to GNSS-challenged environments, has ushered in unprecedented opportunities for developments, and it has become the focus of research due to its advantages of good autonomy, rich visibility, small size and low cost (Scaramuzza et al., 2011; Fuentes-Pacheco et al., 2012; Nistér et al., 2006; Lin et al., 2019). However, with unpredictable texture structure and motion blur continuously exist in mobile platform visual imagery and seriously reduce the similarity between images, accurate, stable and well-distributed matches are difficult to obtain, thus real-time VO/VSLAM and largescale structure from motion pose severe challenges to limited computing resources (Chen et al., 2019).

To address this problem, appropriate key-frame selection strategy can help increase the accuracy and consistency of local motion estimation of VO/VSLAM.

At present, many scholars have make a series of studies on key-frame selection of VO/VSLAM, and put forward many typical solutions (Klein et al., 2007; Tan et al., 2013; Qin et al., 2018; Lin et al., 2018; Wolf., 1996; Zhuang et al., 1998; Mur-Artal et al., 2015; Mur-Artal et al., 2017). It can be classified into the following categories:

(1). Select key-frame according to a fixed time or frame interval. Like parallel tracking and mapping (PTAM) (Klein et al., 2007) requires a high quality with tracking result when selecting key-frames. The selected key-frame needs to meet an exact transformation and rotation angle. The shortest distance between the camera and the nearest key point of the map makes it difficult to triangulate the new feature points.

(2). Insert key-frames according to the image overlap. For example, robust monocular SLAM in dynamic environments (RD-SLAM) (Tan et al., 2013) needs to meet the following three conditions to select key-frames when facing each new frame: a. the camera position and attitude can be successfully estimated; b. the number of feature points extracted in the current frame should exceed a specific threshold; c. shared feature points in existing key-frames are less than a specific threshold.

(3). Insert key-frames according to parallax. It is well known that the VINS-Mono (Qin et al., 2018; Lin et al., 2018) has two requirements when selecting key-frames: a. average parallax, if the average parallax of tracking feature exceeds a certain threshold, the frame will be regarded as a key-frame; b. tracking quality, if the number of tracking features is less than a certain threshold, the frame will be selected as a key-frame.

(4). Insert key-frame according to the image content index (Wolf., 1996; Zhuang et al., 1998). This type of method first establish the feature clustering space of the current frames, then the feature distance between the current frame and the next frame is calculated, and the key frame is selected by the feature distance threshold. Its calculation efficiency is high, but its accuracy is difficult to guarantee.

(5). Other type. The famous ORB-SLAM (Mur-Artal et al., 2015) and ORB-SLAM2 (Mur-Artal et al., 2017) adopt the key-frame selection strategy as a survival of the fittest by inserting key-frames as quickly as possible, which can remove later redundant

* Corresponding author

frames, achieve robustness in difficult scenarios and avoid additional costs.

Although numerous researches have been done on key-frame selection of VO/VSLAM, some methods that select key-frame at equal distances or intervals at the same amount of time may have lack flexibility, other key-frame selection methods, such as image overlap, parallax, content indexing, marginalization and information entropy may consume a lot of time to repeat feature extraction, to match and to calculate.

In order to make a balance between the flexibility and the real-time performance of key-frame selection strategy, we propose a simple and efficient way which is different from the current method. Our method based on the essential matrix estimation and improved PWC-Net (Niklaus et al., 2018).

Figure 1. The prediction result of improved PWC-Net. The first and second images were taken by camera. The third image is an intermediate image predicted by the net.

As shown in Figure.1, the improved PWC-Net is a context-aware synthesis approach that warps not only the input frames but also their pixel-wise contextual information and uses them to interpolate a high-quality intermediate frame. The method first use a pre-trained neural network to extract per-pixel contextual information for input frames, then it employ a state-of-the-art optical flow algorithm to estimate bidirectional flow between them and pre-warp both input frames and their context maps. Finally, the method feeds the pre-warped frames and their context maps to a video frame synthesis neural network to produce the interpolated frame in a context-aware fashion.

This algorithm predicts the motion state of the corresponding mobile platform by the change of attitude between frames in a certain interval. If the attitude change between frames exceeds the given threshold, then the improved PWC-Net is used to associate the image content of the corresponding front and back frames, and then the intermediate frame is filled for the odometry as the key-frame. This makes the key-frames denser in complex cases, and relatively sparse in flat areas.

The major contributions of this study can be summarized as follows:

1. In order to verify the feasibility of the method, the pyramid layered KLT tracking (Lucas et al., 1981), five-point method (Nister et al., 2004) and RANSAC (Nister et al., 2005) algorithm are used to calculated the inter-frame attitude by two-step decomposition of essential matrix;

2. The motion state of the vehicle is predicted by the change of inter-frame attitude within a certain interval. The improved PWC-Net is automatically used to add a middle frame between two key-frames when the estimated attitude fails to meet the requirements, which enables densely selection of key-frame in the complex situation, and sparsely in the flat area;

3. The effectiveness of the method is verified by public dataset as KITTI.

2. MRTHODOLOGY

2.1 The overall architecture

For all the mobile platforms with cameras, the slighter the attitude of the platforms change, the higher the image overlap rate will be, so a sparse frame rate can meet the accuracy requirements of platforms attitude estimation. However, a sharp turn, uphill, downhill or a lateral shaking can easily force the attitude of platforms to change greatly, so a higher frame rate is significantly required to help improving the estimate of the platforms' motion state. As shown in Figure 2, the attitude of a mobile platform will generally face several situations: a. yaw (heading) angle changes around the Y axis when the platform moving along the horizontal plane; b. pitch angle changes around the X axis when the platform moving uphill and downhill; c. roll angle changes around the Z axis when the platform facing lateral jitter occurs; d. combinations of three cases.

In the article, we through two-step decomposition to get the change of posture angle between frames (essential matrix →rotation matrix →attitude angle) to predict the motion statue of mobile platforms in a certain time interval. If the attitude change between frames do not exceeds the given threshold, the key-frames will be selected in certain interval. If the attitude change between frames exceeds the given threshold, then the improved PWC-net is used to associate the image content of the corresponding front and back frames.

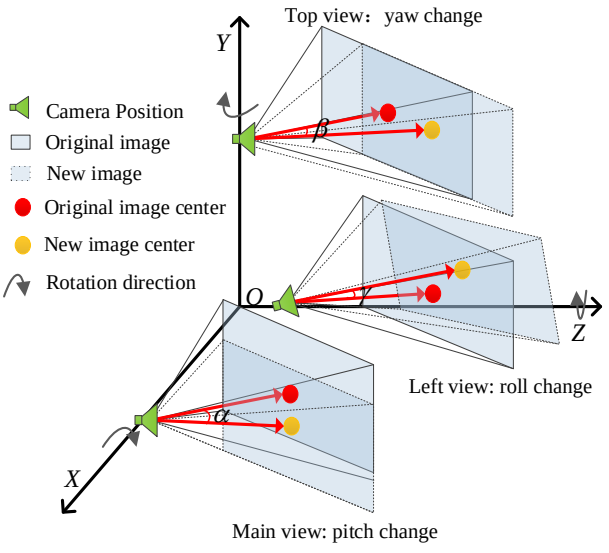


Figure 2. Schematic diagram of attitude change between frames

In order to select key-frames when the local motion of camera is consistent within a short time interval, we propose a key-frame select and artificial make algorithm according to the change of attitude angle. The proposed method can be summarized in Algorithm 1 and Figure 3.

Algorithm 1 An Automatic Key-Frame Selection Method for Visual Odometry.

Input: Sequence images or videos.
Output: local motion state estimation of camera, key-frames sequence F.
1: Read sequence images $f_1, f_2, \dots, f_i$ or videos and preprocessing.
2: Initialize the key-frames sequence F: the first frame image and the second frame image are separately stored into F, and tracking the next frame, if fails, the adjacent two frames are sequentially selected and stored into F.
3: KLT tracking: for new frame $f_i, i > 3$, using FAST algorithm to detect feature points in f_i, then tracking these feature points in f_{i+1}; if the number of feature points tracked is less than a certain threshold, redetect the feature points in f_i, and tracking corresponding feature points in f_{i+1} to obtain corresponding feature point pairs between frames.
4: For the feature point pairs in step3, Nister's five-point method and the RANSAC algorithm are used to calculate the essential matrix E_i.
5: The essential matrix E_i is decomposed into a rotation matrix R_i and a translation vector t_i.
6: Determine whether the rotation matrix R_i is nonsingular and the translation vector t_i is small, if not, return to step 3.
7: Decompose the rotation matrix R_i into pitch angle α along the X axis, heading angle β along the Y axis, and roll angle γ along the Z axis.
8: If $\alpha < m_\alpha || \beta < m_\beta || \gamma < m_\gamma$, then $F \leftarrow f_{net} \leftarrow f_i$, where $m_\alpha, m_\beta, m_\gamma$ are thresholds of attitude angle change, F is the key-frame sequence, f_{net} is the key-frame that made by the improved PWC-net between f_{i-1} and f_i, $i = 1,2, \dots, n$ (n is the number of frames), go to step 3; otherwise $F \leftarrow f_i \leftarrow f_i^k$, ($k$ is the number of maximum interval frames, less than half of the frame rate in this paper),where $k = 1$ and $i = 1,2, \dots, n$, go to step 3.
9: Return local motion state estimation of camera and key-frame sequence F.

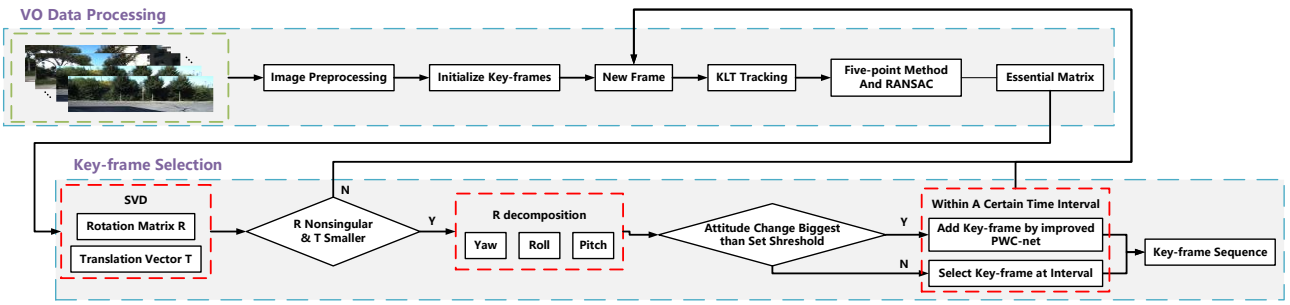


Figure 3. The flow of key-frame selection method

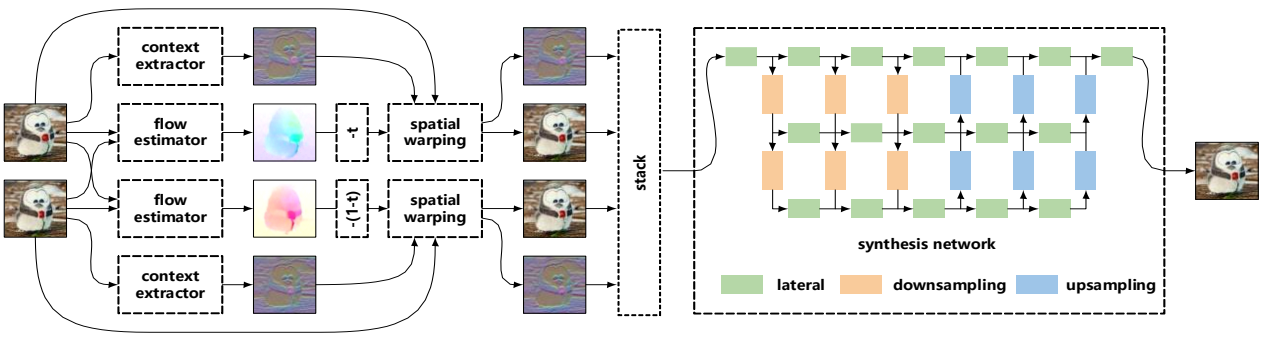


Figure 4. The flow of the improved PWC-Net

2.2 Two-step decomposition

If the two sets of same image coordinate points corresponding to frames F_p and F_q in two given space as $P = \{p_1, p_2, \dots, p_n\}$ and $Q = \{q_1, q_2, \dots, q_n\}$, respectively, and the frame F_p coincides with the frame F_q after rotation and translation of the elements with external orientation (R, t), which can be expressed as:

$$Q = RP + t \tag{1}$$

Where $R = \begin{bmatrix} r_{00} & r_{01} & r_{02} \\ r_{10} & r_{11} & r_{12} \\ r_{20} & r_{21} & r_{22} \end{bmatrix}$, $R * R^T = I$, $\det(R) = 1$.

This group of Euler angles, pitch angle, yaw angle and row angle, can describe the motion state of the mobile platform, also known as the attitude angle. Assuming that the attitude angle of the platform rotating on three coordinate axes X, Y and Z is divided as pitch angle α, yaw angle β and roll angle γ, the calculation formula of direction cosine matrix (DCM) R is as follows:

$$R(\alpha, \beta, \gamma) = R_z(\gamma) R_y(\beta) R_x(\alpha)$$

$$= \begin{bmatrix} c_\beta c_\gamma & s_\alpha s_\beta c_\gamma - c_\alpha s_\gamma & s_\alpha s_\gamma + c_\alpha s_\beta c_\gamma \\ c_\beta s_\gamma & c_\alpha c_\gamma + s_\alpha s_\beta s_\gamma & c_\alpha s_\beta s_\gamma - s_\alpha c_\gamma \\ -s_\beta & s_\alpha c_\beta & c_\alpha c_\beta \end{bmatrix} \tag{2}$$

The pitch angle and roll angle cannot be divided after the heading angle is determined, so it can be assumed that these two angles are influenced and determined by each other. Generally, if the roll angle is assumed to be zero, the attitude angle can be expressed as follows:

$$\begin{cases} \alpha = arctan(-r_{12}, r_{11}) \\ \beta = arcsin(-r_{20}) \\ \gamma = 0 \end{cases} \quad (3)$$

2.3 Improved PWC-Net (Niklaus et al., 2018)

The goal of the improved PWC-Net is to generate an intermediate frame $\hat{f}_t$ between the given two consecutive video frames f_1 and f_2. The improved PWC-Net runs in three stages that is illustrated in Figure 4. The method first estimates the bidirectional optical flow between f_1 and f_2 and extracts the pixel-level context map. Then the input frame and its context mapping are distorted according to the optical flow. The method final inputs them into a deep frame to synthesize neural networks to generate intermediate frame image.

The method estimates the bidirectional optical flow $F_{1\to2}$ and $F_{2\to1}$ between the two frames using the recent PWC-Net method (Sun et al., 2018), which combines warpage and cost volumes. The method also uses multi-scale feature pyramid, so it performs well in standard benchmark test and has high calculation efficiency.

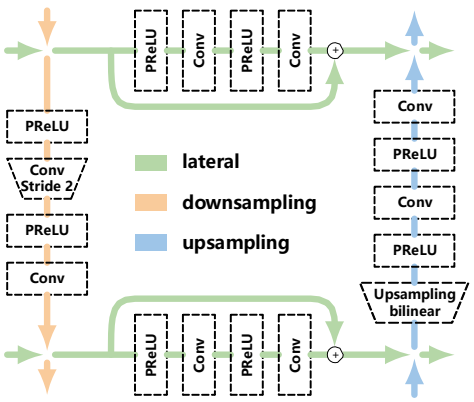


Figure 5. Building block of the frame synthesis neural network in Figure 4, adapted from the GridNet architecture.

As shown in Figure 5, the method modifies the horizontal and vertical connections, so a GridNet can learns how information at different scales should be combined on its own, making it well-suited for pixel-wise problems where global low-resolution information guides local high-resolution predictions.

3. EXPERIMENTAL RESULTS AND ANALYZES

In this section, description of datasets, implementation details and experimental results are provided.

3.1 Description of datasets

KITTI datasets: In order to evaluate the performance of our key-frame selection method proposed in this paper, the KITTI dataset experiments were carried out respectively. The data collection environment incudes urban areas and suburbs.

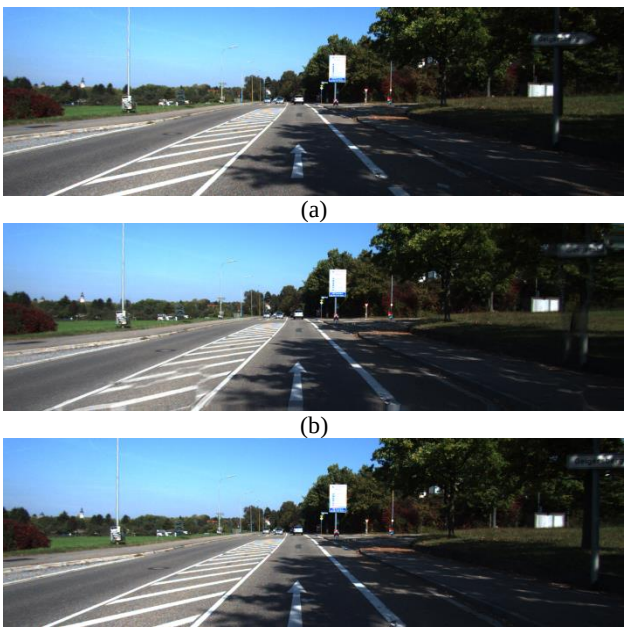
Figure 6. Examples of city and suburb scene in the KITTI dataset

3.2 Implementation details

In terms of threshold setting, in order to ensure that the KLT tracking is not lost and the motion state of the vehicle is accurately recovered, the threshold of the inter-frame attitude angle change and the interval step size should be smaller. The change of the inter-frame attitude angle and the interval step size are usually determined by the running speed of the vehicle and the camera frame rate. Therefore, they are used as the basis for selecting the threshold of attitude angle change and the interval step size when performing key-frame selection. We implemented our approach using PyTorch with improved PWC-Net. We have carried out our experiments with an Intel® Core™ i9-8950HK (2.90GHz) and 32GB RAM.

3.3 Experimental results and analyses

3.3.1 Analysis of the intermediate frames make by improved PWC-Net.



(a) (b) (c)

Figure 7. The experimental result of road scene. (a) and (c) are the input frames captured by KITTI. (b) is the intermediate frame make by improved PWC-Net.

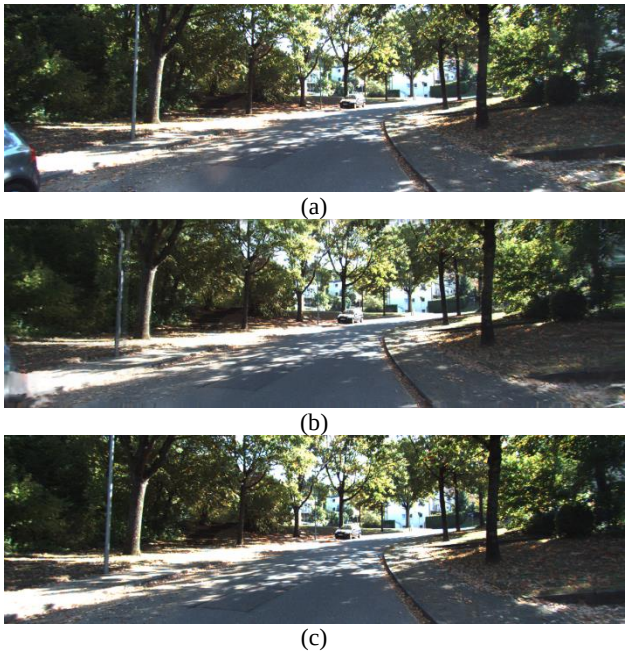


Figure 8. The experimental result of woodland scene. (a) and (c) are the input frames captured by KITTI. (b) is the intermediate frame make by improved PWC-Net.

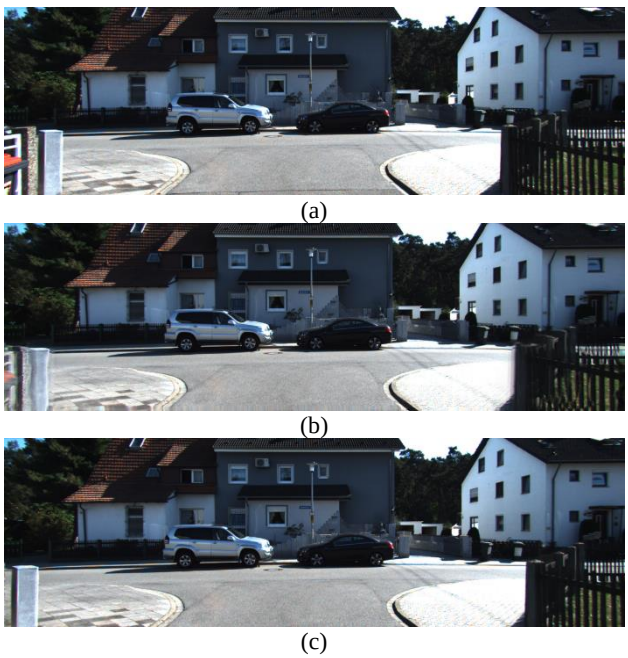


Figure 9. The experimental result of town scene. (a) and (c) are the input frames captured by KITTI. (b) is the intermediate frame make by improved PWC-Net.

Figures 7 to 9 show the effect of the selected scenario on the network's frame complement results under the improved PWC-Net architecture. As shown in Figures 7-9, whether it is an open road, a wooded forest, or a town with a large number of buildings, the improved PWC-Net is very robust facing input frames containing different scenes.

However, the PWC-Net still has some shortcomings when it automatically generating intermediate frames. For example, in Figure 7, on the far right side of the generated image (b), the tree is virtual, this is because when the sensor is acquiring images, due to the running speed or dynamic vibration of the mobile platform, or the significant change of the object shooting angle, the acquired image has visible object deformation at the edge of the image (especially the left and right sides of the image). The deformation of the object makes the network unable to grasp the specific shape and position of the corresponding object when it generates intermediate image. This situation is more obvious when the edge objects of the current image disappear in the next frame, such as the car on the left of Figure 8 and the stone pillar on the left of Figure 9. So in the future research, we will focus on the improved PWC-Net structure, hoping to solve or improve the corresponding problems.

3.3.2 Analysis of visual odometery trajectory prediction results

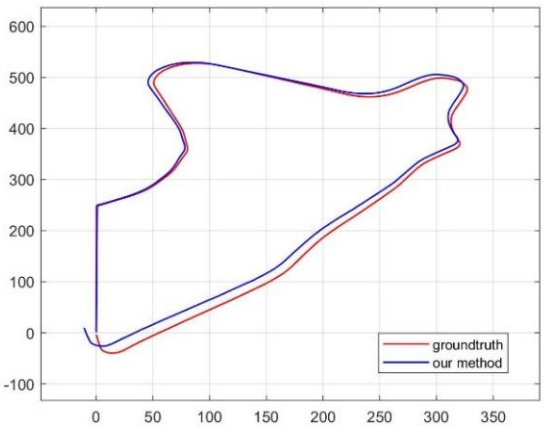


Figure 10. The experimental result of full trajectory of KITTI datasets.

As shown in Figure 10, the result of the improved PWC-Net based key-frame selection strategy are compared with the ground-truth. The red trajectory is made according to the groundtruth and the blue trajectory is the running result of our algorithm. As can be seen from the Figure 10, the trajectory shape of the experimental results in this paper is complete, and the radian prediction when turning basically conforms to the actual trajectory given by the true value. The blue trajectory did not form an ideal closed loop in the end, mainly due to the accumulation of errors generated during each turn. However, it can be seen from the figure that the algorithm of this paper is extremely accurate when the moving receipt is driving in a straight line.

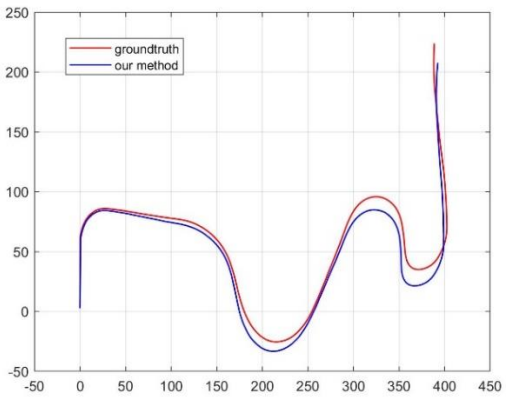


Figure 11. The experimental result in the multi curve scene.

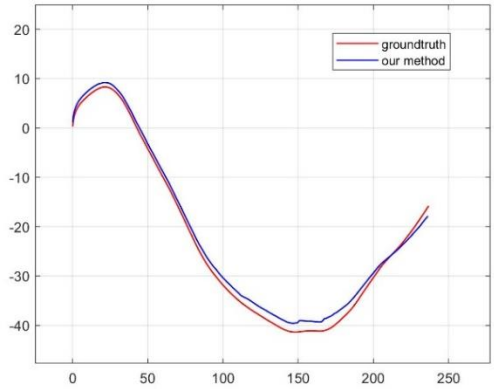


Figure 12. The experimental result in the large curve scene.

As shown in Figure 11 and Figure 12, they are the trajectory prediction results made by our key-frame selection strategy when facing turning. It can be seen from the figures that the experimental results of the algorithm in this paper are acceptable at the turning compared with the groundtruth trajectory. This is because our strategy realizes the dense filling of key-frames to a large extent when the field of vision changes greatly in the face of large-scale or multiple turns.

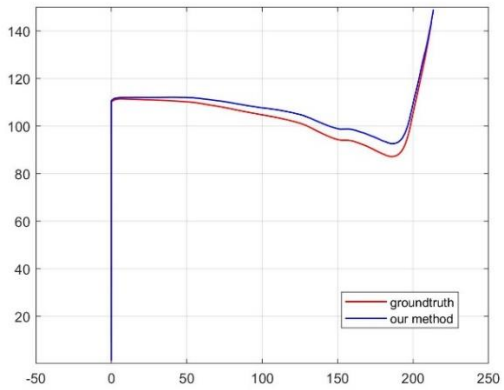


Figure 13. The experimental result in the straight scene.

As can be seen in Figure 13, the algorithm in this paper can achieve better racking result when the mobile platform moving straight or facing right angle turn. However, the slight error after turning will always produce a large error accumulation after the long-distance straight-line movement. To some extent, the offset error of the algorithm in this paper is due to the use of optical flow method visual odometer. In the absence of inertial navigation information, the optical flow algorithm is difficult to have robust performance in the face of the change of light brightness and the lack of brightness feature points in a certain range.

We plan to replace the optical flow visual odometer with the feature point visual odometer in the future research, and plan to focus the matching on the central area of each image.

3.3.3 Comparison of key-frame selection method with the method without key-frame selection.

From Figure 10 to Figure 13, it can be seen that the key-frame selection method can achieve better accuracy both in global and local enlargement. Correspondingly, Table 1 shows the total frames (TF), the number of key-frames (NKF) and time consuming of the proposed method with key-frame selection (KFS) and without KFS.

Methods	TF	NKF	Time(s)
With KFS	1590	1099	280.378
Without KFS	1590	1587	354.492

Table 1. Comparison of key-frame selection method with the method without key-frame selection

As can be seen from Table 1, the number of key-frames and time consuming of the proposed method with key-frame selection are greatly reduced compared to the method without key-frame selection. Proper key-frame selection can increase the accuracy of feature point triangulation, and then improve the local accuracy.

4. CONCLUSIONS AND FUTURE DIRECTIONS

This paper proposed an automatic key-frame selection method based on the improved PWC-Net for monocular visual odometry. In order to verify the proposed method, we conducted extensive experiments with KITTI datasets in complex scenarios. The results showed that we achieved sparsely selection of key-frame at the straight line area and densely at the sharp turn area. To evaluate the proposed method, from qualitative aspects: the effectiveness of the key-frame selection method is intuitively displayed from the comparison of the key-frame trajectories with different steps and reference trajectories in the global and local enlargement. From quantitative aspects: the relationship between the number of key-frames and time consuming of the experiment with different thresholds were counted. We draw the conclusion that the proposed method can greatly reduce data redundancy and improves the real-time performance of VO/VSLAM with relatively high accuracy.

However, due to the influence of Covid-19 and the closure of Wuhan, many comparative experiments and data acquisition of this method are limited. We hope to improve the key-frame selection and filling strategy after returning to Wuhan University, change the structure of the reference neural network, and compare with the existing algorithm, so as to further verify the feasibility and effectiveness of the algorithm proposed in this paper.

ACKNOWLEDGEMENTS

We would like to thank anonymous editors and reviewers for their kind suggestions. This study is supported by The National Key Research and Development Program of China under grant no. 2017YFC0803802.

REFERENCES

D. Scaramuzza and F. Fraundorfer, “Tutorial: Visual odometry,” *IEEE Robot. Autom. Mag.*, vol. 18, no. 4, pp. 80-92, Dec. 2011. doi.org/10.1109/MRA.2011.943233.

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, Nov. 2012. doi.org /10.1007/s10462-012-9365-8.

D. Nistér, O. Naroditsky, and J. Bergen, "Visual odometry for ground vehicle applications," *J. Field Robot., vol.* 23, no. 4, pp. 3-20, Jan. 2006.

X. Lin, F. Wang, L. Guo, and W. Zhang, "An Automatic Key-Frame Selection Method for Monocular Visual Odometry of Ground Vehicle," J. IEEE Access, vol. 7, pp. 70742-70754, 2019. doi.org/10.1109/ACCESS.2019.2916901.

Y. Chen, L. Yan, B. Xu, and Y. Liu, "Multi-Stage Matching Approach for Mobile Platform Visual Imagery," J. IEEE Access, vol. 7, pp. 160523-160535, 2019. doi.org/10.1109/ACCESS. 2019.2950909.

G. Klein and D. Murray, "Parallel tracking and mapping for small AR workspaces," in *Proc. 6th IEEE ACM Int. Symp. Mixed Augment. Reality*, Nov. 2007, pp. 1-10.

W. Tan, H. Liu, Z. Dong, G. Zhang, and H. Bao, "Robust monocular SLAM in dynamic environments," in *Proc. IEEE Int. Symp. Mixed Augment. Reality (ISMAR)*, Adelaide, SA, Australia, Oct. 2013, pp. 209-218.

T. Qin, P. Li, and S. Shen, "VINS-mono: A robust and versatile monocular visual-inertial state estimator," *IEEE Trans. Robot.*, vol. 34, no. 4, pp. 1004-1020, Aug. 2018.

Y. Lin, F. Gao, T. Qin, W. Gao, T. Liu, W. Wu, Z. Yang, and S. Shen, "Autonomous aerial navigation using monocular visual-inertial fusion," *J. Field Robot.*, vol. 35, no. 1, pp. 23-51, Jan. 2018. doi.org/10.1002/rob.21732.

W. Wolf, "Key frame selection by motion analysis," in *Proc. IEEE Int. Conf. Acoust.*, Atlanta, GA, USA, May 1996, pp. 1228-1230.

Y.T. Zhuang, Y. Rui, T. S. Huang, and S. Mehrotra, "Adaptive key frame extraction using unsupervised clustering," in *Proc. IEEE Int. Conf. Image*, Chicago, IL, USA, Oct. 1998, pp. 866-869.

R. Mur-Artal, J. M. M. Montiel, and J. D. Tardós, "ORB-SLAM: A versatile and accurate monocular SLAM system," *IEEE Trans. Robot.*, vol. 31, no. 5, pp. 1147-1163, Oct. 2015. doi.org /10.1109 /TRO.2015.2463671

Mur-Artal, Raul, and Juan D. Tardós. "Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras." *IEEE Trans. Robot.* 33.5 (2017): 1255-1262. doi.org /10.1109/TRO.2017.2705103

Niklaus S, Liu F. "Context-aware synthesis for video frame interpolation"*Proc. IEEE Conf. Computer Vision and Pattern Recognition*. 2018: 1701-1710.

B.D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proc. 7th Int. Jt. Conf. Artif. Intell.*, Apr. 1981, pp.121-130.

D. Nister, "An efficient solution to the five-point relative pose problem," IEEE trans. Pat. Anal. Mac. Intel., vol. 26, no. 6, pp. 756-777, Jun. 2004.

D. Nister, "Preemptive RANSAC for Live Structure and Motion Estimation," Mac. Vis. Appl., 2005 16: 321. Vol. 16, no. 5, pp. 321-329, Dec. 2005.

Sun, Deqing, et al. "Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume." *Proc. IEEE. Conf. Computer Vision and Pattern Recognition*. 2018.