A NOVEL GEOMETRIC KEY-FRAME SELECTION METHOD FOR VISUAL-INERTIAL SLAM AND ODOMETRY SYSTEMS

A. Azimi [1], A. Hosseininaveh [1], F. Remondino[*] [2]

[1] Department of Photogrammetry and Remote Sensing, Faculty of Geodesy and Geomatics Engineering, Toosi University of Technology, K. N, Tehran, Iran - arashazimi0032@gmail.com, hosseininaveh@kntu.ac.ir

[2] 3D Optical Metrology (3DOM) unit, Bruno Kessler Foundation (FBK), Trento, Italy - remondino@fbk.eu

KEY WORDS: Visual Odometry, Visual SLAM, Visual-Inertial Systems, IMU, Geometric Key-Frame Selection

ABSTRACT

Given the importance of key-frame selection in determining the positioning accuracy of Simultaneous Localization And Mapping (SLAM) and Odometry algorithms, and the urgent need in this field for a flexible key-frame selection algorithm, this paper proposes a novel and geometric method for key-frame selection built on top of ORB-SLAM3. It takes a key-frame in a completely robust and flexible way regardless of the environment, data and scene conditions, and according to the physics and geometry of the environment. In the proposed method, the camera sensor and IMU take key-frames simultaneously and in parallel. While selecting a key-frame, an adaptive threshold first decides whether the geometric condition of the frame is appropriate based on the degree of change in the orientation of the point visibility vector from the last key-frame to the current frame. Then the quality of the frame is evaluated by examining the distribution of points inside the frame by a balance criterion. A new key-frame will be created if both conditions provide a positive answer. In addition, if the IMU sensor detects large changes in acceleration, a key-frame independently chosen. The proposed method is evaluated qualitatively and quantitatively on the EuRoC dataset by comparing the algorithm trajectory to a reference trajectory and usig the Absolute Trajectory Error (ATE) and the processing time as metrics. The evaluation results indicate a 26% improvement in the positioning of the algorithm although it has a 9% increase in the processing time due to its geometric key-frame selection process.

1. INTRODUCTION

Because of recent advances in robotics and autonomous vehicle research, having an accurate and real-time positioning and mapping technology has become increasingly vital. As a result, due to their benefits of lightweight, cheap cost, low power consumption, and compact size, camera-based systems such as Visual Odometry (VO) and Visual Simultaneous Localization And Mapping (VSLAM) have been noted by many researchers as an excellent supplement to GPS-challenged situations (Fuentes-Pacheco et al. 2015), (Nistér et al., 2006). Many algorithms have been created in this sector, with the ORB-SLAM3 method being the first. It exceeds all prior pioneering algorithms in terms of accuracy, speed, and resilience, including SVO (Forster et al., 2014), VINS mono (Qin et al. 2018), DSO (Engel et al. 2017), OKVIS (Leutenegger et al. 2015), etc. However, such algorithms face computational complexity and real-time processing issues as a result of the vast volume of data. Processing only a few important frames rather than all of them is the typical way for resolving this problem and removing data redundancy, which decreases computing complexity while retaining accuracy and consistency. As a consequence, choosing the right key-frames can help VO/VSLAM algorithms become more accurate and consistent.

To select key-frames, heuristic and non-geometric thresholds with limited flexibility are geerally used. In this work we present an efficient key-frame selection approach based on ORB-SLAM3 that substitutes most of the heuristic thresholds with a geometric-based IND-inspired (Hosseininaveh et al. 2012) method. In the proposed method, two geometric criteria are investigated at the same time: 1) an adaptive threshold determines whether or not this frame is appropriate for becoming a key-frame after categorizing the angles between the camera to point vectors and map points surface normal in four 10-degree zones and comparing them to the equivalent zones in the previous key-frame; 2) a balance criteria checks the suitability of the distribution of points inside the frame by calculating the center of gravity of the points inside the frame. In addition, if the IMU sensor detects a significant acceleration change, it generates a new key-frame on its own. The performance of the proposed method are validated by experiments on two image sequences from the EuRoC dataset (Burri et al. 2016). The suggested method presents an efficient and resilient geometric solution for VO / VSLAM key-frame selection, which may be used instead of current heuristic methods.

2. RELATED WORK

In different domains, such as computer vision algorithm, video summarization, photogrammetry and structure from motion (SFM), there has been a substantial amount of study on key-frame selection.

One of the less investigated disciplines is key-frame selection using a deep neural network. Lu Sheng et al. (2019), for example, created a deep network that simultaneously learns key-frame selection and visual odometry tasks. Their studies clearly illustrate the method's usefulness, but they were very dependent on how key-frames in training data were picked.

Video synopsis is one of the video summarization applications for surveillance cameras in order to achieve efficient video browsing and retrieval (Baskurt and Samet 2019, Yan et al. 2020).order to achieve efficient video browsing and retrieval Video processing applications, which are generally not real-time and aim to extract frames containing relevant information from all frames, constitute the next area of key-frame selection study. Zhuang et al. (1998) split the frames into numerous categories based on texture and form, then score each frame using a weighted combination of these features, capturing the key-frame with the highest score in each category. This method has had good results in cases with severe light changes. Besiris et al. (2007) choose

[*] Corresponding author

key-frames based on the notion of greatest frame separation, after categorizing frames using a graph. Wolf (1996) proposed a key-frame selection method based on motion analysis for identifying key-frames in shots from video programs. They use optical flow computations to identify local minima of motion in a shot. This technique allows to identify both gestures which are emphasized by momentary pauses and camera motion which links together several distinct images in a single shot. Their results show that this method can well extract key-frames from a complex shot.

Photogrammetric and 3D reconstruction applications are one of the most significant study areas in the field of key-frame selection. Hosseininaveh et al., (2021) and Ahmadabadian et al. (2013) proposed the image network designer (IND) approach for extracting ideal subsets of images from a sequence of images acquired from an object. In this method, the angle between the normal to the surface in each point and the viewing vector of each point, in each image, classified in four different areas. The camera that covers the most areas of the all points is then selected as the best camera. The findings demonstrate that this technique produces a full and accurate point cloud, as well as a final reconstructed model, with excellent outcomes. The only problem with this method is the inability to run in real-time applications (Hosseininaveh et al. 2014; Hosseininaveh and Remondino 2021). Dong et al. (2014) developed an offline key-frame selection technique that consisted of two parts: an off-line module for selecting features from a set of reference pictures and an online module for matching them to the input live video for estimating the camera posture rapidly (Dong et al., 2014).

Computer vision applications and VO/VSLAM algorithms is one of the most common uses of key-frame selection methods. The biggest difference between these methods and the methods of the previous categories is the ability to run in real-time. These approaches involve visual information such as scene light flow, pixel grays, and so on, as well as positional information such as the distance between frames and the positioning of map points in the key-frame selection process. Engel et al. (2017) first select many key-frames and quickly sparsify them by marginalizing redundant key-frames. To select a key-frame, they introduce a combination of three criteria, focusing on drastic changes in light and scene brightness and the gray-scale value of the pixels. Experiments on numerous datasets have shown that this approach of picking key-frames provides improved outcomes in poor illumination circumstances. Position-based methods in computer vision applications and VO / VSLAM algorithms can be divided into the sub-categories including based on 1) specified time or place intervals, 2) image overlap, 3) parallax, and 4) others (Lin et al. 2019). Key-frame selection in parallel tracking and mapping (PTAM) (Klein and Murray, 2007), Semi-direct monocular visual odometry (SVO) (Forster et al. 2014), and Large-scale direct monocular SLAM (LSD-SLAM) (Engel et al. 2014) are based on the first category and without considering any specific criteria and only with the passage of a particular time or distance intervals key-frames are selected. OKVIS (Leutenegger et al. 2015) and SLAM in dynamic environments (RD-SLAM) (Tan et al. 2013) exploit the image overlap methods, the second category, and have more flexibility and more power than the methods using the previous category criterion. VINS-mono (Qin et al. 2018), as an instance in the third category, has two criteria for selecting key-frames including average parallax and tracking quality. An example for the last category is Kerl et al. (2013) who presented a key-frame selection method based on differential entropy of multivariate normal distribution that had excellent results in texture-less environments but it is computationally complex.

Xiaohu lin et al., (Lin et al., 2019) select key-frames based on the relative variations of the Roll, Pitch, Yaw angles. If camera attitude changes sharply, the key-frame selection rate increases, and if camera attitude shifts slightly, Key-frames are taken at a lower rate. The results show that this method increases the algorithm's speed by reducing 40% - 60% of the redundant frames and, at the same time, does not reduce the positioning accuracy. Finally, the ORB-SLAM algorithm (Mur-Artal et al., 2015), by adopting the best key-frame selection strategy, first takes a lot of key-frames and then marginalizes their redundancies to maintain the algorithm's performance. ORB-SLAM3 select a key-frame if four requirements are met: 1) It must have been more than 20 frames since the last global re-localization. 2) Local mapping is inactive, or there has been more than 20 frames since the previous key-frame insertion. 3) At least 50 points are tracked in the current frame. 4) The current frame hasn't tracked more than 90% of the points from the previous key-frame. Experiments have shown that the ORB-SLAM3 method is strong and reliable, and that it has delivered good results in difficult circumstances.

3. THE PROPOSED METHOD

3.1 Method overview

The proposed method is based on ORB-SLAM3 (Campos et al. 2020) and is aimed to improve the accuracy and robustness of the visual SLAM algorithm. The suggested method uses a key-frames selection methodology based on geometric and photogrammetric concepts, as well as adaptive (rather than static) thresholds to the greatest extent possible. Furthermore, employing the synchronized IMU sensor and camera, this technique utilizes a key-frame. The camera sensor uses geometric and photogrammetric principles inspired by the IND approach to determine whether or not to pick a key-frame (Ahmadabadian et al. 2013; Hosseininaveh et al. 2012). Simultaneously, the IMU sensor picks the current frame as the key-frame if it detects considerable acceleration changes. Figure 1 shows the method's schematic diagram.

3.2 Camera key-frame selection

Two geometric criteria participate in the selection of the key-frame by the camera: 1) adaptive threshold and 2) balance criteria. The adaptive threshold is explained first. For each new input frame, there are a number of points that are also present in the last key-frame. After categorizing the angle of visibility vector of each of these points relative to the normal vector on the surface, in four 10-degree zones, the points whose visibility vector zone has changed from the last key-frame so far are counted.

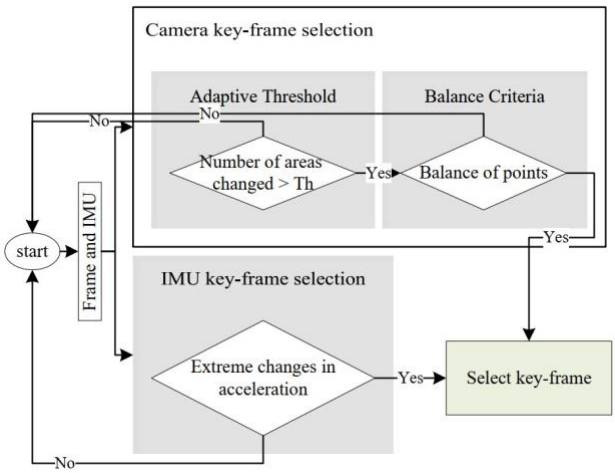

Figure 1. The flowchart of the proposed key-frame selection method.

The role of the adaptive threshold is to control the number of these

changes, and if these changes exceed the adaptive threshold, a key-frame will be allowed by the adaptive threshold. The 10 degree angle of the conical zones has been selected following the investigations of Hosseininaveh et al. (2012).

To define this adaptive threshold, we need the most similar frame to the last key-frame, called the reference frame. The frame that enters immediately after the key-frame is considered the reference frame.

Firstly, an initial threshold is estimated to calculate the adaptive threshold. This initial value, assuming that the current frame and the reference frame are similar, is selected in such a way that the ratio of the points whose area has changed to the total corresponding points is equal in the reference frame and the current frame.

This initial threshold is simplistic and has to be modified because this frame is not identical to the reference frame and has been moved and changed. By the ratio of decreasing the number of matched points from the reference frame to the current frame, a coefficient is employed to make the initial threshold tougher. By adding this coefficient, the adaptive threshold is strict and does not allow key-frames; Because it considers the change in the area of the points visibility vector only due to the decrease in the number of corresponding points, which is the result of the displacement of the frame; While these changes may be due to poor lighting conditions and so on. Therefore, another coefficient is considered which simplifies the initial threshold by decreasing the ratio of points whose area has changed to all corresponding matched points.

After applying the mentioned coefficients, the initial threshold is fully adapted and can be adapted to any situation; But the remarkable thing about this threshold is that it does not pay attention to the quality of the current frame to become a key-frame. As a result, to check the quality of the frame, the balance criterion of the points inside the frame is activated and the distribution of points inside the frame is examined.

To calculate the balance criterion, a 3-by-3 grid is first created inside each frame, and inside each cell of this grid, the number of points whose area has changed are counted. This process creates a 3-by-3 matrix for each image. By calculating the center of gravity of this matrix (Johnson, 2013) for all frames that satisfy the condition of the adaptive threshold, the frame whose center of gravity is closer to the center of the matrix is selected as the key-frame. Satisfaction of these two criteria gives us the assurance that the selected key-frame, in addition to having a good geometric condition, its quality is also suitable for matching and pairing with the previous key-frame.

3.3 IMU key-frame Selection

It is possible to track the quick and abrupt motions in which the camera fails in visual-inertial systems, owing to the IMU sensor, and increase the algorithm's stability in this condition. The IMU is utilized to pick the key-frame in abrupt movements in the method presented in this work, so that the moment of rapid movement can be detected and a key-frame can be acquired to avoid the algorithm from failing.

The key-frame may be selected by IMU using a simple threshold since the acceleration values are absolute and independent of data and frame state. The experimentally determined threshold is 1 (meter/second^2). A key-frame is adopted if the acceleration surpasses this threshold, and this procedure is independent of the camera sensor's key-frame selection mechanism.

4. EXPERIMENTS AND RESULTS

The EuRoC Micro Aerial Vehicles (MAV) dataset (Burri et al., 2016) was used to test the performance of the key-frame selection method proposed in this work. Stereo images, synchronized IMU measurements, and precise motion and structural ground-truth are all included in the datasets. The proposed method and ORB-SLAM3 were then evaluated quantitatively and qualitatively by comparing the algorithm's trajectory to the ground truth trajectory, as well as the Absolute Trajectory Error (ATE) (Sturm et al., 2012) for two image sequences from EuRoC dataset. The processing time of the algorithms was also compared. All experiments were performed with an Intel (R) Core i7- 4510U (4 cores @ 2 GHz) and 8 GB of RAM. Each dataset was run 10 times and the average was utilized to eliminate some unpredictability in the findings.

4.1 Data and material

The EuRoC dataset is one of the most widely used datasets for evaluating computer vision algorithms in automated navigation scenarios (Burri et al., 2016). There are 11 image sequences including simple, medium and difficult level flights in this dataset, which includes accurate ground truth position information measured by laser scanners and IMU information synced with frames. The ORB-SLAM3 and the algorithm presented in this paper were evaluated in mono-inertial and stereo-inertial modes for two image sequences from this dataset (MH01, MH02) and their trajectories are compared with the reference trajectories, also the value of Absolute Trajectory Error (ATE) and the processing time of the algorithms is obtained for them. The calculated trajectory and the ground truth are aligned using a similarity transformation (Zhang and Scaramuzza, 2018) to determine the ATE.

4.2 Comparison of the trajectories

Experiments were carried out on two sequences of the EuRoC dataset (MH01, MH02) in stereo-inertial and mono-inertial modes to qualitatively validate the key-frame selection method proposed in this study. Figures 2 and 3 illustrate the trajectories compared to the reference trajectories in stereo-inertial and mono-inertial modes, respectively.

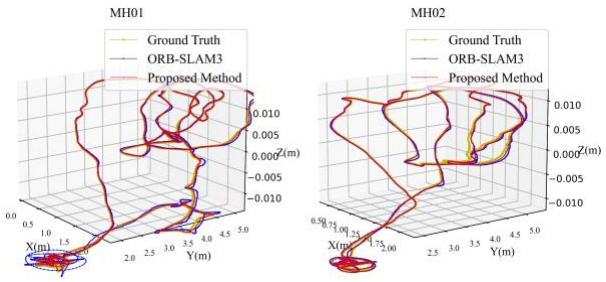

Figure 2. Frame trajectories in the stereo-inertial mode.

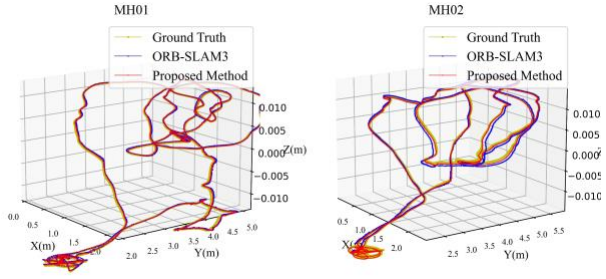

Figure 3. Frame trajectories in the mono-inertial mode.

Figures 2 and 3 indicate that the proposed method overtook ORB-SLAM3 in terms of performance and trajectory deviation. The

divergence between the ORB-SLAM3 and the reference trajectory has grown as the route turns. The method given in this study, on the other hand, has retained its closeness to the reference trajectory.

4.3 Comparison of the Absolute Trajectory Error - ATE

The ATE is used to determine positioning accuracy in order to quantitatively assess the suggested approach. This criteria was computed ten times in two image sequences for stereo-inertial and mono-inertial modes of both algorithms, and the average of the results can be seen in Table 1. Figures 4 and 5 illustrate the cumulative ATE values of ten time runs for stereo-inertial and mono-inertial modes, respectively.

	Stereo-Inertial		Mono-Inertial	
Sequence	*ORB-SLAM3*	*Proposed Method*	*ORB-SLAM3*	*Proposed Method*
MH01	0.0366	0.0310	0.0350	0.0185
MH02	0.0318	0.0280	0.0625	0.0410
Total	0.0342	0.0295	0.0488	0.0298

Table 1. Average ATE results of two algorithms applied to the two sequences. Unit: [m].

Each number in the table represents the average of ten times the execution of each algorithm in each image sequence, with the Total row representing the average of both data. The improvement of the accuracy of the algorithm proposed in this paper compared to the ORB-SLAM3 is evident from the table above.

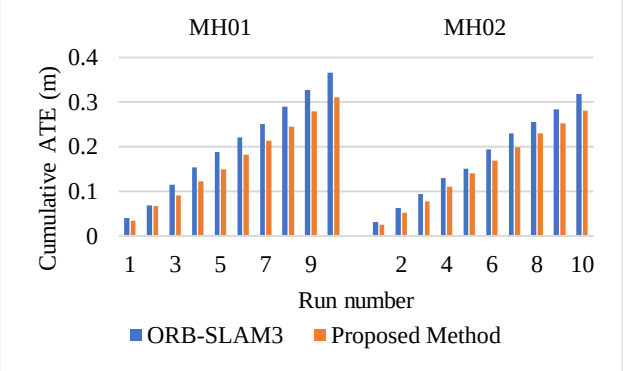

Figure 4. Cumulative ATE results in stereo-inertial mode.

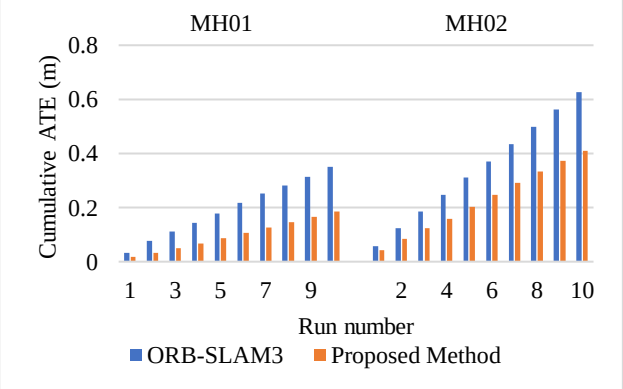

Figure 5. Cumulative ATE results in mono-inertial mode.

The significantly decrease in the amount of ATE with increasing number of runs is evident from the image above.
Figure 6 also displays the average value of each algorithm's ATE outputs in these two sequences.
The results of Table 1 and Figures 6 and 7 show an improvement of 13.7% in stereo-inertial mode and 38.9% in mono-inertial mode of the algorithm presented in this paper. Figure 7, which is the average ATE of both data for each algorithm, also shows the reduction of the ATE difference between mono-inertial and stereo-inertial modes for the proposed algorithm. Reducing the difference between ATE in mono-inertial and stereo-inertial mode in the algorithm proposed in this paper, indicates more stability and less effectiveness of this algorithm from the type of system used.

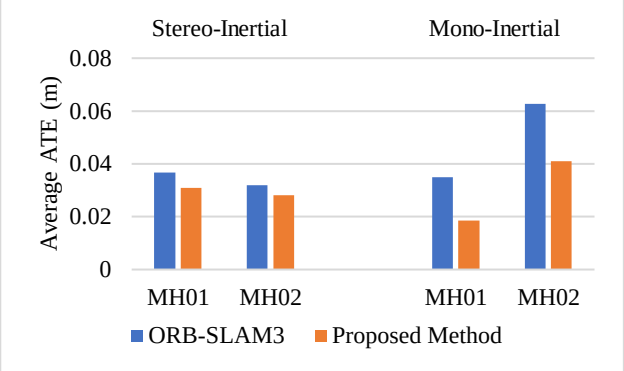

Figure 6. Average of ATE values in the two dataset.

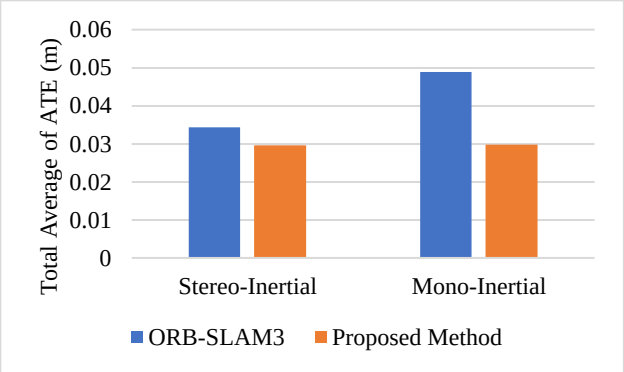

Figure 7. Total average of ATE values in stereo-inertial and mono-inertial modes.

4.4 Comparison of processing time

The processing time of the method proposed in this paper is expected to rise compared to ORB-SLAM3 due to its geometric key-frame selection process. As a result, the processing time of each method is measured in this section. To determine the processing time, each algorithm is run for both image sequences and the processing time is recorded. The time measurement accuracy in this evaluation is 0.1 seconds. The results are given in the table 2.

Sequence	*ORB-SLAM3 (sec)*	*Proposed Method (sec)*
MH01	350.0	379.1
MH02	280.5	306.2

Table 2. Processing time of each algorithm for both sequence.

The execution time of the algorithm presented in this paper is higher than ORB-SLAM3, as expected, due to the computational

complexity of obtaining the key-frame in this algorithm. Because the ORB-SLAM3 algorithm obtains key-frames using innovative and fixed thresholds, whereas the proposed algorithm obtains key-frames using geometric constraints, an almost constant value is added to the algorithm processing time per frame; as a result, the proposed algorithm imposes more complex calculations on the algorithm while improving positioning accuracy. However, these additional calculations do not have a tangible effect on the real-time execution of the algorithm and the algorithm is still executed in real-time.

5. DISCUSSIONS

As the selection of key-frames is the foundation of positioning in SLAM and Odometry algorithms, the precision and manner of chosing these frames will have a substantial influence on the algorithm's accuracy – and successive 3D reconstruction tasks. Due to the use of predefined thresholds specifically fine-tuned for standard data, existing key-frame selection algorithms do not work effectively in diverse settings and in non-standard data, despite their high accuracy in standard data (such as the EuRoC dataset). As a result, an attempt has been made in this article to introduce a flexible geometric method for picking key-frames that is efficient in all scenarios. There are a few key considerations to consider regarding this method: as mentioned in Section 3.2, the angle of the cone zones according to (Hosseininaveh et al. 2012; Ahmadabadian et al., 2013) is considered to be 10-degrees. Small camera motions cause point zones to alter and key-frames to be picked faster when this angle is reduced. This increases computing time while also weakening the intersecting triangle's geometry. Increased this angle, on the other hand, decreases the key-frame selection rate and hence the key-frame network's stability. Optimal selection of this parameter will help to improve the positioning accuracy of the algorithm. Another issue worth mentioning is the acceleration change threshold used by the IMU to choose a key-frame. This threshold is set based on the camera's mounting platform. It will be larger on faster-moving flying platforms and smaller on slower-moving ground platforms. In this study, the value of this threshold is set at 1 (meter/second^2) by experimentation.

6. CONCLUSIONS

This paper proposed a novel geometric key-frame selection method for visual-inertial SLAM and Odometry systems built on ORB-SLAM3 framework. Extensive tests with two sequences from the EuRoC dataset in mono-inertial and stereo-inertial modes were conducted to assess the proposed method. The results demonstrated that we were able to create a completely geometric key-frame selection procedure that worked reliably and consistently in a variety of settings without the need of heuristic thresholds. By comparing the algorithm trajectory to the reference trajectory and the ATE, our approach was assessed quantitatively and qualitatively. The proposed algorithm shows a 25-30% improvement in accuracy, although the processing time is slightly longer, requiring some further optmizations for real-time processing operations.

In future research, the proposed algorithm might be modified to choose key-frames in such a manner that a dense and coherent point cloud is produced, in addition to further enhance positioning accuracy. Our method may be used as a basic algorithm in the generation of training data for deep learning networks, and its speed can be enhanced with the help of deep learning networks.

REFERENCES

Ahmadabadian, A. H., S. Robson, J. Boehm and M. Shortis, 2013. Image selection in photogrammetric multi-view stereo methods for metric and complete 3D reconstruction. Proc. of *Videometrics, Range Imaging, and Applications XII; and Automated Visual Inspection*. International Society for Optics and Photonics.

Baskurt, K. B., Samet, R., 2019. Video synopsis: A survey. *Computer Vision and Image Understanding*, 181: 26-38.

Besiris, D., Laskaris, N., Fotopoulou, F., Economou, G., 2007. Key frame extraction in video sequences: a vantage points approach. Proc. *IEEE 9th Workshop on Multimedia Signal Processing*.

Burri, M., Nikolic, J., Gohl, P., Schneider, T., Rehder, J., Omari, S., Achtelik, M. W., Siegwart, R., 2016. The EuRoC micro aerial vehicle datasets. *International Journal of Robotics Research*, 35(10): 1157-1163.

Campos, C., Elvira, R., Rodríguez, J. J. G., Montiel, J. M., Tardós, J. D., 2020. ORB-SLAM3: An accurate open-source library for visual, visual-inertial and multi-map SLAM. *arXiv preprint arXiv*:2007.11898.

Dong, Z., Zhang, G., Jia, J., Bao, H., 2014. Efficient keyframe-based real-time camera tracking. *Computer Vision and Image Understanding*, 118: 97-110.

Engel, J., Koltun, V. Cremers, D., 2017. Direct sparse odometry. IEEE *Transactions on PAMI*, 40(3): 611-625.

Engel, J., Schöps, T., Cremers, D., 2014. LSD-SLAM: Large-scale direct monocular SLAM. Proc. *ECCV*.

Forster, C., Pizzoli, M., Scaramuzza, D., 2014. SVO: Fast semi-direct monocular visual odometry. Proc. *IEEE ICRA*.

Fuentes-Pacheco, J., Ruiz-Ascencio, J., Rendón-Mancha, J.M., 2015. Visual simultaneous localization and mapping: a survey. *Artificial Intelligence Review*, 43(1): 55-81.

Hosseininaveh, A. and Remondino, F., 2021. An Imaging Network design for UGV-based 3D reconstruction of buildings. *Remote Sensing*, 13(10): 1923.

Hosseininaveh, A., Sargeant, B., Erfani, T., Robson, S., Shortis, M., Hess, M., Boehm, J., 2014. Towards fully automatic reliable 3D acquisition: from designing imaging network to a complete and accurate point cloud. *Robotics and Autonomous Systems*, 62(8): 1197-1207.

Hosseininaveh, A., Serpico, M., Robson, S., Hess, M., Boehm, J., Pridden, I., Amati, G., 2012. Automatic image selection in photogrammetric multi-view stereo methods. Proc. 13th *VAST Symposium*, Eurographics Association.

Johnson, R. A., 2013. *Advanced euclidean geometry*. Courier Corporation.

Kerl, C., Sturm, J. Cremers, D., 2013. Dense visual SLAM for RGB-D cameras. Proc. *IEEE/RSJ International Conference on Intelligent Robots and Systems*.

Klein, G. and Murray, D., 2007. Parallel tracking and mapping for small AR workspaces. Proc. *6th IEEE ISMAR*.

Leutenegger, S., Lynen, S., Bosse, M., Siegwart, R., Furgale, P. 2015. Keyframe-based visual–inertial odometry using nonlinear optimization. *International Journal of Robotics Research*, 34(3): 314-334.

Lin, X., Wang, F., Guo, L., Zhang, W., 2019. An automatic key-frame selection method for monocular visual odometry of ground vehicle. *IEEE Access*, 7: 70742-70754.

Mur-Artal, R., Montiel, J.M.M., Tardos, J.D., 2015. ORB-SLAM: a versatile and accurate monocular SLAM system. IEEE *Transactions on Robotics*, 31(5): 1147-1163.

Nistér, D., Naroditsky, O., Bergen, J., 2006. Visual odometry for ground vehicle applications. *Journal of Field Robotics*, 23(1): 3-20.

Qin, T., Li, P., Shen, S., 2018. Vins-mono: A robust and versatile monocular visual-inertial state estimator. IEEE *Transactions on Robotics*, 34(4): 1004-1020.

Sheng, L., Xu, D., Ouyang, W., Wang, X. 2019. Unsupervised collaborative learning of keyframe detection and visual odometry towards monocular deep slam. Proc. *IEEE ICCV*.

Sturm, J., Engelhard, N., Endres, F., Burgard, W., Cremers, D., 2012. A benchmark for the evaluation of RGB-D SLAM systems. *IEEE/RSJ International Conference on Intelligent Robots and Systems*.

Tan, W., Liu, H., Dong, Z., Zhang, G., Bao, H., 2013. Robust monocular SLAM in dynamic environments. Proc. *IEEE ISMAR*.

Wolf, W., 1996. Key frame selection by motion analysis. Proc. *IEEE International Conference On Acoustics, Speech, And Signal Processing.*

Yan, X., Gong, H., Jiang, Y., Xia, S.-T., Zheng, F., You, X., Shao, L., 2020. Video scene parsing: an overview of deep learning methods and datasets. *Computer Vision and Image Understanding*, 201: 103077.

Zhang, Z. and Scaramuzza, D., 2018. A tutorial on quantitative trajectory evaluation for visual (-inertial) odometry. Proc. *IEEE/RSJ IROS*.

Zhuang, Y., Rui, Y., Huang, T. S., Mehrotra, S., 1998. Adaptive key frame extraction using unsupervised clustering. Proc. *ICIP*.