A Novel Adaptive Keyframe Selection Method with Multi-source Joint Constraints for Visual SLAM

Hongmei Chen, Henan University of Technology, Zhengzhou, China

Baocun Wang, Henan University of Technology, Zhengzhou, China

Dongbing Gu, University of Essex, Colchester, UK

Wen Ye, National Institute of Metrology, Beijing, China

Accepted for publication in Intelligent Service Robotics.

A Novel Adaptive Keyframe Selection Method with Multi-source Joint Constraints for Visual SLAM

Abstract

The effectiveness of keyframe selection is crucial for visual Simultaneous Localization and Mapping (SLAM) algorithms. Stability and adaptability of the selection are challenging issues in complex environments. This paper introduces a novel Adaptive Keyframe Selection method with Multi-Source Joint Constraints (MSJCA-KS), which addresses the challenges by exploiting camera geometry, real-time inertial measurement unit (IMU) data analysis, and effective point distribution constraints. Particularly, we used adaptive threshold conditions based on camera geometry for the keyframe selection under varying scenes. A real-time state detection mechanism is used to refine these thresholds by using IMU data during rapid motion. Furthermore, a constraint mechanism is used to ensure a balanced distribution of keyframes. Our keyframe selection method is integrated into the ORB-SLAM3 framwork to perform SLAM tasks. The effectiveness of our method is validated on the EuRoC dataset and the TUM dataset in terms of Absolute Trajectory Error (ATE), keyframe quality, and computational efficiency. When compared to the Photogrammetric Keyframe Selection (PKS) algorithm, the proposed method achieves a 16% improvement in localization accuracy in the EuRoC dataset and 27% and 23% in the TUM dataset for monocular and stereo cameras, respectively.

Keywords: Visual SLAM, keyframe selection, multi-source joint constraints, adaptive threshold, IMU data, localization accuracy

1 Introduction

Simultaneous Localization and Mapping (SLAM) technology has emerged as a crucial research topic in modern robotics and intelligent industries [1, 2]. Visual SLAM is particularly popular in intelligent perception for mobile robots due to its real-time capability, cost-effectiveness, and ability to capture color texture information [3]. By treating map points and poses from all frames as independent variables in nonlinear optimization, the cumulative error of the front-end pose is minimized, leading to improved positioning accuracy in visual SLAM [4]. However, optimizing visual SLAM poses challenges in terms of computational complexity and real-time requirements,

primarily due to the substantial amount of data involved [5]. To address this issue, it is essential to reduce data redundancy while maintaining accuracy and reliability [6]. The selection of appropriate keyframes plays a vital role in improving the accuracy and real-time performance of visual SLAM algorithms [7].

Existing keyframe selection methods often rely on heuristic thresholds, which exhibit limited adaptability and precision in complex environments due to their simplistic constraint conditions [8]. To overcome the limitations, the paper proposes a new Multi-Source Joint Constraint Adaptive Keyframe Selection(MSJCA-KS) method, which is then integrated into the ORB-SLAM3 framwork to perform SLAM tasks.

Our method makes three key contributions. Firstly, it utilizes camera geometry to define a four-region spatial cone for keyframe selection, which could improve the adaptability by tailoring the selection process to varying scenes. Secondly, a state detection mechanism is used to activate the IMU fusion constraints, which could adapt to rapid motion in complex environments and accordingly improve the robustness of our method. Lastly, a uniform distribution of effective points is used to ensures the stability of keyframe selection.

The remaining structure of this paper is outlined as follows: section 2 presents a review of the related works on keyframe selection of visual SLAM. Section 3 introduces the methods of the keyframe selection algorithm proposed in this paper, discussing the rationale behind the adaptive threshold conditions, state detection mechanism, and effective point distribution criteria. Section 4 provides a detailed description of implementation of our method. Finally, section 5 summarizes the main contributions of this paper.

2 Related works

The keyframe selection methods for visual SLAM are mainly classified into four categories: keyframe selection methods based on motion patterns, data association, appearance, and hybrid approaches [9]. In addition, some researchers considered the use of deep neural networks for keyframe selection [10]. However, due to their heavy dependence on training data, these approaches have received less attention in the computer vision domain [11].

For motion based selection methods, PTAM [12], SVO [13], LSD-SLAM [14], RGB-D SLAM [15], etc., employed specific time interval thresholds or spatial interval thresholds for keyframe selection. The RANSAC method in FAST-SLAM [16] introduced a keyframe selection approach based on average image pixel displacement, utilizing a static threshold to control the displacement. Thormählen et al. [17] used the Geometric Robust Information Criterion (GRIC) to reduce frame position estimation errors. However, this criterion is applicable only to the initialization of frame positions and cannot be used for tracking. OKVIS [18] and RD-SLAM [19] leveraged the information of image overlap for keyframe selection. VINS-mono [20] proposed two criteria, average disparity and tracking quality, for keyframe selection. Kerl [21] and others employed the differential entropy of the multivariate normal distribution to calculate the entropy between the motion estimates of last keyframe and current

frame. This method performs well in textureless environments but is computationally complex, and insufficient lighting can reduce its efficiency. In [22], keyframes were chosen by considering the relative changes in camera's orientation angles. Significant pose changes increase the keyframe selection rate, while slight changes decrease the keyframe selection rate. This method reduces 40-60% of redundant frames.

Data association based methods have also been explored. Fanfani et al. [23] proposed a keyframe selection approach based on high temporal disparity feature ratio. Frames with large temporal disparities and corresponding uncertainties in three-dimensional positions are selected as keyframes. Stalbaum et al. proposed a method to maximize mutually visible features among multiple keyframes [24]. By employing a keyframe window, this method selects keyframes to enhance long-term feature consistency. While it surpasses spatial or temporal threshold-based approaches in robustness, the reliance on static thresholds for feature matching could limit algorithm generalization in certain scenarios.

For appearance based keyframe selection methods, Engel et al. introduced a method that incorporates variables such as gaze abruptness, camera offset, and exposure time to determine keyframes [25]. The Loop Closure Detection for Sparse Odometry (LDSO) method was developed to enhance the repeatability of feature points while upholding the robustness of DSO [26]. Similarly, the Dense Scene Model (DSM) approach was devised to establish a persistent map that manages re-observations and diminishes redundancy within the system [27]. Nevertheless, these approaches tend to overlook geometric constraints. In response to this gap, Dong et al. introduced an adaptive keyframe selection technique [28] that aims to encompass the entire spatial domain as a geometric constraint, while simultaneously minimizing content redundancy. However, this strategy mandates extensive offline training procedures.

Hybrid keyframe selection methods combine different approaches. Chen et al. [29] proposed utilizing the change in view and its rate between the current frame and the most recent keyframe for keyframe selection. The change in view is estimated by using a dynamic threshold designed by a proportional-differential controller. As this method combines dynamic thresholds and geometric methods for keyframe extraction, it exhibits good localization accuracy. However, the number of keyframes selected by this method is relatively small, limiting the construction to a sparse map. Wang et al. [30] proposed an intelligent hybrid keyframe extraction algorithm by combining background difference method and SIFT feature matching algorithm. ORB-SLAM3 [31] employs a survival-of-the-fittest keyframe selection approach. The tracking thread relaxes the filtering conditions, while the local mapping thread eliminates redundant keyframes to reduce computation. Experimental results demonstrate the robustness and effectiveness of the ORB-SLAM3 algorithm. Azimi et al. [9] introduced a photogrammetric key-frame selection method (PKS) that enhances localization accuracy without compromising real-time performance of the algorithm. Nonetheless, when confronted with scenarios involving intense motion in intricate environments, the keyframe selection process is susceptible to challenges such as missed or multiple selections.

In summary, the motion pattern-based keyframe selection method is straightforward and user-friendly but suffers from low adaptability and accuracy due to its

reliance on static thresholds. This paper presents a method that combines multi-source constraints from camera geometry and IMU measurement to adaptively adjust the selection threshold, significantly enhancing stability and performance in complex scenes. The data association-based approach improves feature point consistency; however, its dependence on static thresholds limits generalization in dynamic environments. We reduce this dependency and enhance robustness through spatial cone point and uniform distribution constraints. Appearance-based methods work well in illumination changes or sparse textures but often overlook geometric constraints, making them less effective for rapid movements. Our approach integrates geometric information, adaptive thresholding, and joint multi-source constraints to maintain stable system performance in variable environments. While the hybrid method merges geometric considerations with static thresholds, it remains less adaptable during rapid motion or significant scene changes. In this paper, we dynamically adjust thresholds via real-time state detection to better handle abrupt motions while improving keyframe selection accuracy and system robustness.

3 Multi-Source Joint Constraint Adaptive Keyframe Selection (MSJCA-KS)

This paper presents a novel keyframe selection method for visual SLAM, which leverages multi-source fusion technology and photogrammetric techniques. The proposed method, MSJCA-KS, addresses the problem of redundancy in video frames by incorporating multiple constraints and joint constraints from various sources [32]. This approach improves the adaptability and stability of keyframe selection, leading to enhanced performance in Visual SLAM systems [33]. The method utilizes both camera and IMU sensors, employing triple constraints with threshold conditions for keyframe selection. Fig. 1 illustrates the keyframe selection process of MSJCA-KS.

3.1 Keyframe Selection Process

The keyframe selection process in our method involves several steps. Firstly, we process the input camera video data and IMU data using camera geometric constraints. This step includes the recovery of three-dimensional map points and the definition of a four-region spatial cone for these points. These points are known as the spatial cone points. To establish an initial threshold, we make initial assumptions and optimize the calculation.

Secondly, we employ a state detection mechanism to assess the current motion state and determine if it is in a drastic motion state. Two criteria are used to discriminate the state: the angular velocity measurement exceeding 0.35 (rad/s) or the acceleration measurement surpassing $1\,\mathrm{m/s^2}$. If either criterion is met, it indicates a drastic motion state. Based on the criteria, we decide whether to activate camera and IMU joint constraints. If the detected motion is not in a drastic motion state, the camera and IMU joint constraints remain inactive, and only the adaptive threshold calculated from camera geometric constraints is used as the keyframe selection criterion. If the current motion is in a drastic motion state, the multi-source joint constraints are activated. In this case, both camera and IMU constraints are considered for keyframe selection.

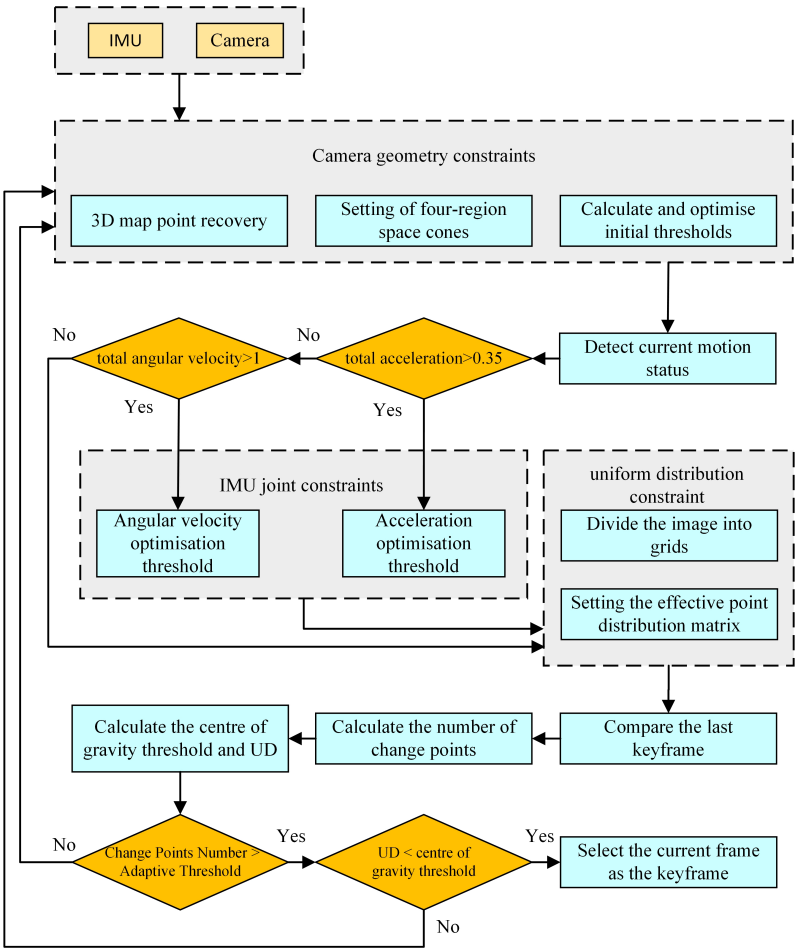

Fig. 1 The keyframe selection process of MSJCA-KS.

Next, we apply effective point uniform distribution constraints. This step involves dividing the image into grids and setting the effective points distribute matrix.

Finally, we compare the current frame with the last keyframe. We calculate the change in the number of spatial cone points the center of gravity threshold, and the Uniform Distribution (UD) criteria. If the change in the number of spatial cone points exceeds the adaptive threshold and the UD is less than the center of gravity threshold, we select the current frame as a keyframe. Otherwise, if these conditions are not met, the current frame is skipped, and the calculation proceeds to the next frame.

3.2 Camera Geometric Constraint

The selection of keyframes in our method is inspired by the Image Network Designer (IND) [34] and Photogrammetric Keyframe Selection (PKS) [9] methods. We employ

adaptive thresholds based on real-time inter-frame matching conditions for keyframe selection [35].

3.2.1 Four-region Spatial Cone

Referring to the PKS method, we first create a Four-region Spatial Cone [9] as shown in Fig.2. The frames are categorized into four groups: the last keyframe, the reference frame, ordinary frames, and the current frame. The four-region spatial cone is centered on one of the map points with its main axis normal to the surface at the point location and its zones separated from the cone's main axis by 10 degree angles. Vectors from matched points to map points are zoned based on their angles to the cone's axis. Frame movement alters these angles, prompting an adaptive threshold to determine key-frame selection, based on the count of points with shifted sight lines. The points with shifted sight lines are also called the changed spatial cone points.

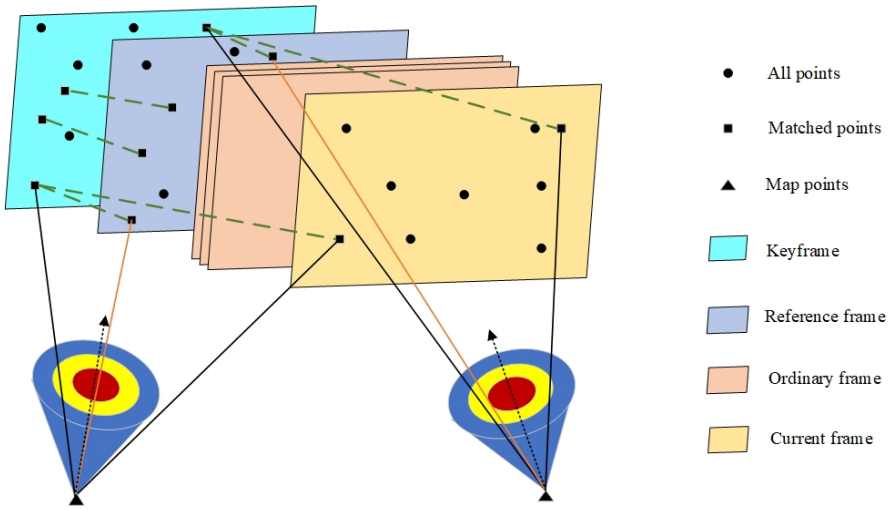

Fig. 2 Four-region spatial cone

3.2.2 Initial Thresholds

Upon system initialization, only a limited number of video frames are available. It becomes necessary to establish an initial threshold for keyframe selection. Refer to the PKS initialization method [9],the system assumes that the current frame's state is similar to the reference frame. When the current frame is similar to the reference frame, the number of feature points, the number of tracking points and the number of changed spatial cone points are nearly identical in the both frames. To calculate the initial threshold, we consider two factors: the number of tracked points and the number of feature points. To balance the influence of all the extracted feature points and tracked matches from both frames, the average value is taken as the initial threshold to ensure the accuracy of the threshold.

The initial threshold is defined as in equation (1):

$$T_i = \frac{(E_{c1} + E_{c2})}{2}. \tag{1}$$

where T_i represents the initial value of the threshold. E_{c1} and E_{c2} show the number of points in the current frame which their cone's zone has changed from the last key-frame.The definitions of E_{c1} and E_{c2} are given as in equation (2):

$$E_{c1} \cong \frac{D_c}{D_r} E_r, E_{c2} \cong \frac{B_c}{B_r} E_r. \tag{2}$$

E_r represents the number of points in the reference frame where changes occurred in the cone's zone compared to the last keyframe. D_c represents the total number of feature points in the current frame, D_r represents the total number of feature points in the reference frame. Similarly, B_c represents the number of points tracked in the current frame compared to the last keyframe, and B_r represents the number of points tracked in the reference frame compared to the last keyframe.

3.2.3 Threshold Optimization

This assumes that the current frame is similar to the reference frame. However, in most cases, the current frame and the reference frame are not adjacent and the current frame may differ significantly from the reference frame, directly applying the initial threshold for keyframe selection reduces its effectiveness. In this case, the threshold needs to be adjusted accordingly. To tackle this issue, Inspired by the PKS method [9], our research focuses on optimizing the threshold by assigning coefficients based on the system's geometric relationships, thereby aligning it more accurately with real-world scenarios.

Firstly, to ensure a more reasonable initial threshold, it becomes necessary to reduce the number of matching points between the current frame and the reference frame. This can be achieved by introducing a coefficient in percentage form, denoted as α, which modifies the initial threshold. The coefficient is defined in the following equation (3):

$$\alpha = \frac{B_{\mathrm{r}} - B_c}{B_{\mathrm{r}}}. \tag{3}$$

Secondly, refining keyframe selection by considering various scenarios.

Keyframe selection should take into account the following situations:

(1) If the current frame significantly deviates from the previous keyframe after multiple frames, should the current frame be chosen as a keyframe?

(2) Whether the scene information in the current frame differs significantly from the previous keyframe.

These considerations directly impact the effectiveness and adaptability of keyframe selection. To ensure the quality of selected keyframes, avoid the introduction of excessive redundant information, and prevent tracking loss, this study refines the threshold by introducing a second coefficient, denoted as η, which is multiplied by the initial coefficient α. By adjusting the coefficient η, we simultaneously reduce the probability of selecting frames immediately following the last keyframe as keyframes, as well as the probability of selecting frames five frames after the last keyframe as keyframes. Instead, we increase the likelihood of selecting keyframes approximately five frames after the last keyframe. The coefficient η is defined in the following equation (4):

$$\eta = \frac{5 - \Delta d}{3}. \tag{4}$$

Where Δd is the number of video frames between the current frame and the last keyframe.

After multiple experiments and verification, it was found that the number of points undergoing changes within the cone area of the current frame did not exceed the previously calculated adaptive threshold. Consequently, selecting key frames became impractical. To address this, an additional threshold was introduced to simplify the initial threshold, enabling normal key frame selection. Moreover, considering that over half of the matching points in the current frame experience changes within the cone area, designating the current frame as a key frame is appropriate. Achieving this requires increasing the ratio of change points in the cone area to the total number of matching points. This approach enhances selection criteria stability in terms of geometric relationships and ensures the threshold's strictness.This condition is defined by the third coefficient, denoted as φ, in equation (5):

$$\varphi = \frac{(E_{c1} + E_{c2})}{B_c} - \frac{2E_{\mathrm{r}} + B_{\mathrm{r}}}{2B_{\mathrm{r}}}. \tag{5}$$

Finally, considering all the aforementioned factors, the initial threshold is adjusted by incorporating the first and second coefficients as percentages, while subtracting the third coefficient as a percentage. This modification results in the derivation of the final adaptive threshold, expressed as shown in the equation (6):

$$T_a = T_i + \alpha\eta T_i - \varphi T_i. \tag{6}$$

After optimizing the three coefficients mentioned above, the adaptive threshold is refined to better align with the actual situation, effectively meeting the demands of camera motion in practical scenarios.

3.3 IMU Joint Constraints

When solely considering the geometric constraints of the camera, rapid camera movements can lead to sudden changes in the field of view. In such cases, using the aforementioned adaptive threshold as the condition for key frame selection may not be stringent enough. This can result in a rapid increase or decrease in the number of selected key frames, thereby compromising the overall performance of the SLAM system. An excessive increase in the number of key frames will prolong the computation time, while a significant decrease can lead to reduced accuracy or even tracking failure. To address these challenges posed by complex motion scenarios, additional constraints are introduced by incorporating IMU data.

Fig. 3 illustrates the design of a state detection mechanism. This mechanism identifies instances of drastic motion by detecting an overall angular velocity exceeding 0.35 (rad/s) or an overall acceleration surpassing $1\,\mathrm{m/s^2}$. Upon such detection, the IMU joint constraints are activated in conjunction with the camera geometric constraints. Consequently, the camera and IMU data are combined to update the adaptive thresholds, thereby imposing more stringent criteria for key frame selection to accommodate the sudden motion conditions. Conversely, when the state detection indicates non-drastic motion, the IMU joint constraints are bypassed, and the system proceeds with the constraint of maintaining a uniform distribution of effective points.

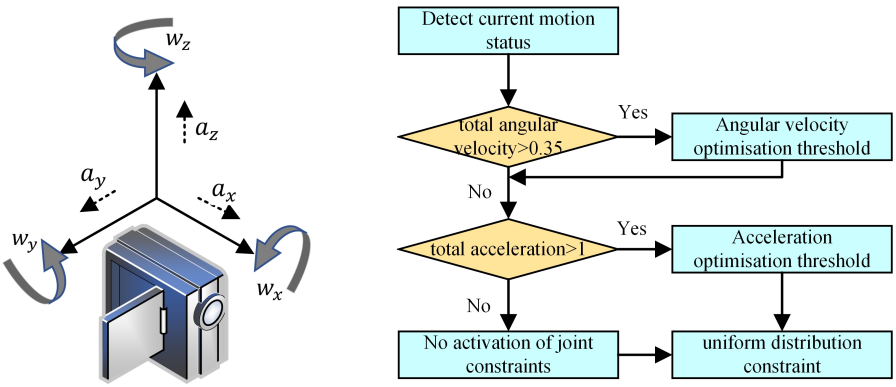

Fig. 3 IMU joint constraint diagram.

In Fig. 3, the left plot displays the average angular velocity and acceleration along the three axes, obtained by pre-integrating IMU measurements. In visual-inertial SLAM systems, the IMU provides measurements of linear acceleration and angular velocity relative to the body coordinate system at the current moment. Integration of these measurements over time yields the average angular velocity and acceleration between frames. To enhance accuracy, numerical processing is applied, considering factors such as zero-bias errors and scale factors. These refinements improve the modeling of physical motion, accounting for measurement and calculation errors in practical systems.

3.3.1 Angular velocity optimization thresold

When the camera viewpoint experiences rapid changes, solely relying on the adaptive threshold derived from camera geometry constraints in equation (6) as the keyframe selection criterion results in a substantial rise in the number of selected keyframes. These surplus redundant keyframes not only fail to improve the system's accuracy but also contribute to heightened computational complexity. To tackle this issue, we introduce a coefficient γ, which is defined as shown in equation (7):

$$\gamma = \frac{1}{1-\omega}. \tag{7}$$

In the equation, ω_T represents the angular velocity measurement in rad/s, which acts as the indicator for rapid angular displacement during intense motion scenarios. Its calculation incorporates the magnitude of the angular velocity of the gyroscope in the IMU in three directions of pitch, yaw and roll; ω_x, ω_y and ω_z are the angular velocities around the x, y and z axes respectively. Taking the Euclidean norm of these three directional angular velocities as the total angular velocity. This is defined as shown in equation (8):

$$\omega_T = \sqrt{\omega_x^2 + \omega_y^2 + \omega_z^2}. \tag{8}$$

Furthermore, we update the coefficient η to adjust the keyframe selection probability. Instead of the original setting of increasing the probability after five frames from the previous keyframe, we now increase it after seven frames. This modification is reflected in the updated equation (9), replacing the original equation (4):

$$\eta = \frac{7-\Delta d}{3}. \tag{9}$$

The updated adaptive threshold is expressed as in equation (10):

$$T_a = \gamma\left(T_i + \alpha\eta T_i - \varphi T_i\right). \tag{10}$$

3.3.2 Acceleration optimization thresold

When the state detection mechanism identifies a rapid acceleration of the camera, a coefficient λ is introduced, denoted as Equation (11):

$$\lambda = \frac{1}{10^a}. \tag{11}$$

The symbol a_T represents the acceleration measurement ($\mathrm{m/s^2}$), which serves as an additional criterion for detecting rapid motion states. Its value considers the magnitude of acceleration in the upward-downward, left-right, and forward-backward directions of the accelerometers in the IMU;a_x, a_y, and a_z are the accelerations along the x, y, and z axes respectively. The Euclidean norm of the acceleration in these three directions is taken as the total acceleration, defined as in equation (12):

$$a_T = \sqrt{a_x^2 + a_y^2 + a_z^2}. \tag{12}$$

In conjunction with this, we apply further updates to the coefficient η to adjust the keyframe selection probability. We now increase it after three frames. This adjustment is aimed at optimizing the keyframe selection process. The updated expression is given by equation (13):

$$\eta = \frac{3 - \Delta d}{3}. \tag{13}$$

To address the issue of a large acceleration a when activating the multi-source joint constraints, we introduce the coefficient λ and update the coefficient η. This aims to reduce the adaptive threshold, facilitating the selection of the current frame as a keyframe. By doing so, we prevent a decrease in localization accuracy or tracking failure caused by an excessive number of keyframes. The updated expression for the adaptive threshold is provided in equation (14):

$$T_a = \lambda \left(T_i + \alpha \eta T_i - \varphi T_i \right). \tag{14}$$

The above-mentioned multi-source joint constraints significantly enhance the adaptability and stability of the system in keyframe selection.

3.4 Uniform Distribution Constraint

To maintain the robustness of the keyframe selection mechanism, it is crucial to consider the variability in the quantity of map points within the four-region space cone, which can affect the alignment between consecutive frames and keyframes. The adoption of a uniform distribution criterion for effective point dispersion is instrumental in enhancing the stability of the selection process. This criterion stipulates that keyframes should be extracted from frames exhibiting a balanced distribution of points. As illustrated in Fig. 4, each frame is subdivided into a nine-cell grid (three rows by three columns)[35]. Points within each cell that manifest a change in the cone region beyond a predefined angular threshold – empirically set at 30 degrees – qualify as effective points, as referenced in [9]. A tally of these effective points is maintained for each cell. Concurrently, this methodology generates a matrix corresponding to the grid for every frame, with each element of the matrix denoting the count of effective points in its

associated cell. This systematic approach ensures that the chosen keyframes reflect a consistent and equitable distribution of points, thereby reinforcing the stability and accuracy of the keyframe selection process.

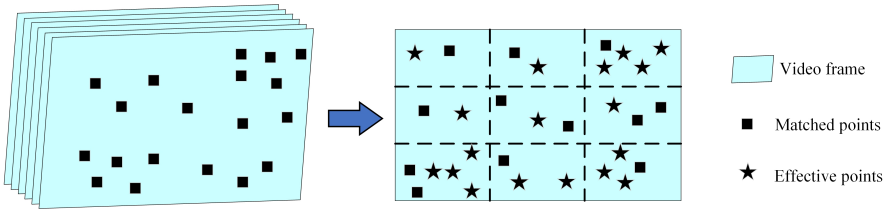

Fig. 4 Generation of the matrix for the distribution of valid points.

In an optimal arrangement, the two largest values within a matrix, MAX_1 and MAX_2, are positioned at the maximum possible separation from each other. To quantify this spatial relationship, we introduce the Uniform Distribution (UD) criterion, which yields a value of 1 when MAX_1 and MAX_2 are at this ideal distance. For any other distribution, the UD value is computed to fall between 0 and 1, reflecting the degree of uniformity in the points' spacing. The initial formulation of the UD criterion, as specified in Equation (15):

$$UD = \frac{\mathrm{di} \times MAX_2}{2\sqrt{2}MAX_1}. \tag{15}$$

considers the relative positions of MAX_1 and MAX_2 within the matrix grid. In this formulation, $2\sqrt{2}$ represents the maximum cell-to-cell distance in the matrix, while di denotes the actual distance between MAX_1 and MAX_2.

In dense data point configurations, traditional uniform distribution criteria fail to provide adequate constraints. This inadequacy is evidenced by the proximity of the matrix's largest (MAX_1) and second-largest (MAX_2) elements, which are often contiguous, yielding a minimal difference (di) compared to the maximum inter-cell distance. This agglomeration of points into a grid-like formation significantly skews the uniform distribution (UD) value, undermining keyframe selection processes. To counter this, an improved uniform distribution criterion is proposed, which includes a discriminant condition: should MAX_1 surpass double the value of MAX_2, the UD formula is recalibrated as presented in Equation (16):

$$UD = \frac{\mathrm{di} \times (MAX_1 - MAX_2)}{2\sqrt{2}MAX_1}. \tag{16}$$

To facilitate this criterion, a threshold is established by computing the matrix's center of gravity within a three-row, three-column grid matrix, following Equation (17):

$$T_h = \sqrt{x_c^2 + y_c^2}. \tag{17}$$

The coordinates x_c and y_c represent the matrix's center of gravity. A UD value falling below this threshold signifies an acceptable distribution of valid points, and thus, the frame is deemed sufficiently stable to qualify as a keyframe.

4 Experiments and results

The experiment utilized the EuRoC micro aerial vehicle (MAV) dataset and the TUM_VI dataset to evaluate the performance of the proposed keyframe selection method [36]. The EuRoC dataset is widely acknowledged as a benchmark for assessing computer vision algorithms in autonomous navigation scenarios. Each sequence in this dataset includes stereo images, synchronized IMU measurements, and ground truth motion trajectories. Similarly, the TUM_VI dataset provides high-quality image sequences, synchronised IMU measurements and precise camera motion trajectories, facilitating the evaluation and comparison of various visual inertial navigation systems. To comprehensively evaluate the proposed method, MSJCA-KS, in comparison to ORB-SLAM3 [31] and PKS [9], the absolute trajectory error (ATE) was calculated by comparing the estimated trajectories with the ground truth trajectories. The evaluations were conducted in both monocular inertial and stereo inertial modes. The experiments were performed on an Intel(R) Core i5-8300H CPU (8 cores @ 2.3GHz) with 16GB RAM. Each dataset was run 10 times, and the average results were used to mitigate random errors in the outcomes.

4.1 Localization Accuracy

Localization accuracy is evaluated by using the root mean square error of the absolute trajectory (RMSE AT) as a metric. Table 1 and Table 2 illustrate that the RMSE AT of the proposed MSJCA-KS is smaller in both monocular and stereo inertial modes compared to the other two algorithms for most sequences. Notably, the performance improvement is particularly obvious in the monocular-inertial mode. Experimental results show that the estimated trajectories by the proposed MSJCA-KS are closely consistent with the ground truth trajectories. In both monocular inertial and stereoscopic inertial modes, the localization accuracy surpasses that of the PKS algorithm by 16% in monocular and stereoscopic modes in the EuRoC dataset. In the monocular inertial mode, the localization accuracy surpasses that of the PKS algorithm by 27%, while in the stereoscopic inertial mode, the positioning accuracy exceeds that of the PKS algorithm by 23% in the TUM_VI dataset.

To qualitatively validate the proposed keyframe selection method, MSJCA-KS, the experiments were conducted on the EuRoC dataset sequences MH02 and MH05 in both monocular and stereo-inertial modes. The estimated trajectories and ground truth trajectories of three algorithms, namely MSJCA-KS, ORB-SLAM3, and PKS, were illustrated as follows: Fig. 5 and Fig. 6 are the projected plane trajectories for the MH02 dataset, Fig. 7 and Fig. 8 are the projected plane trajectories for the MH05

Table 1 The RMSE AT for both monocular-inertial and stereo-inertial modes in the EuRoC (m).

	Mono-Inertial			Stereo-Inertial		
Sequence	ORB-SLAM3	PKS	MSJCA-KS	ORB-SLAM3	PKS	MSJCA-KS
MH01	0.038	0.020	**0.017**	0.025	0.024	**0.019**
MH02	0.068	0.029	**0.018**	0.031	0.039	**0.026**
MH03	0.050	0.033	**0.025**	**0.030**	0.041	0.031
MH04	0.056	0.054	**0.053**	0.051	0.044	**0.044**
MH05	0.083	0.054	**0.046**	0.044	0.043	**0.040**
Total	0.059	0.038	**0.032**	0.036	0.038	**0.032**

Table 2 The RMSE AT for both monocular-inertial and stereo-inertial modes in the TUM_VI (m).

	Mono-Inertial			Stereo-Inertial		
Sequence	ORB-SLAM3	PKS	MSJCA-KS	ORB-SLAM3	PKS	MSJCA-KS
Room1	0.012	0.013	**0.009**	**0.0075**	0.0083	0.0083
Room2	0.044	0.010	**0.008**	0.0062	0.0059	**0.0057**
Room3	0.053	0.015	**0.012**	0.0079	0.0089	**0.0073**
Room4	0.035	0.0094	**0.0089**	**0.0087**	0.0107	0.0088
Slides1	0.672	0.573	**0.456**	0.399	0.439	**0.276**
Slides2	0.595	0.681	**0.462**	**0.185**	0.398	0.357
Total	0.235	0.217	**0.159**	**0.102**	0.145	0.111

dataset, while Fig. 9 and Fig. 10 show the time-varying trajectories for the Room2 and Room3 in the TUM_VI respectively. The comparison demonstrates that the estimated trajectory generated by the proposed MSJCA-KS closely aligns with the actual trajectory and the other three algorithms exhibit deviations from the ground truth. In both turning and straight-line motions, this algorithm exhibits superior performance compared to the ORB-SLAM3 and PKS algorithms.

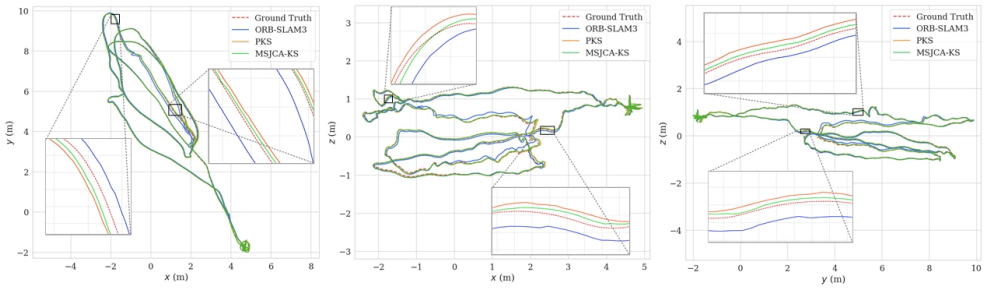

Fig. 5 Comparison of the projected plane trajectories in Monocular-Inertial mode for MH02 in the EuRoC dataset.

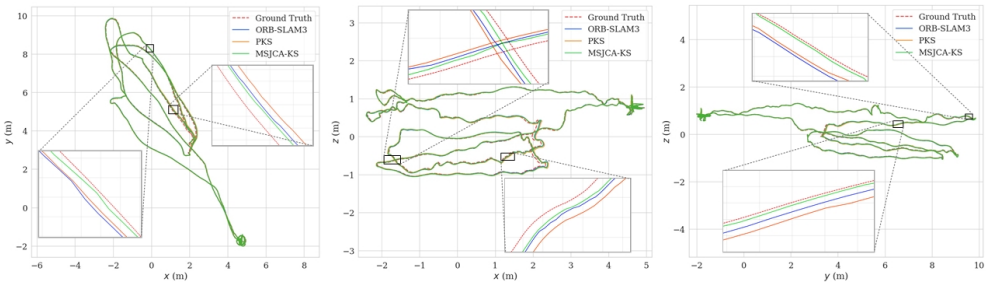

Fig. 6 Comparison of the projected plane trajectories in Stereo-Inertial mode for MH02 in the EuRoC dataset.

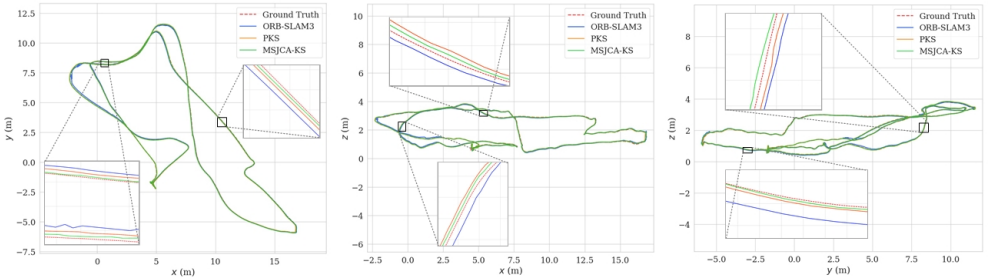

Fig. 7 Comparison of the projected plane trajectories in Monocular -Inertial mode for MH05 in the EuRoC dataset.

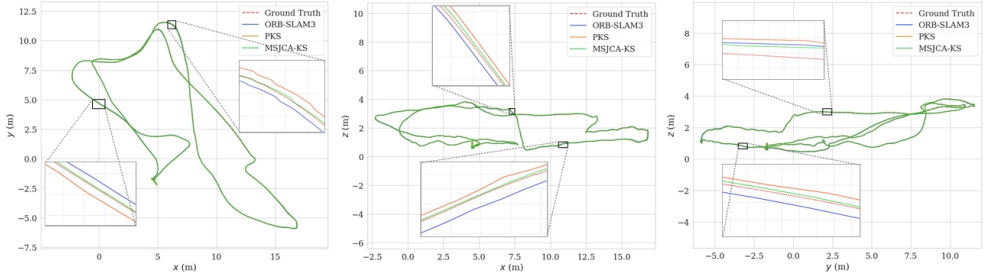

Fig. 8 Comparison of the projected plane trajectories in Stereo-Inertial mode for MH05 in the EuRoC dataset.

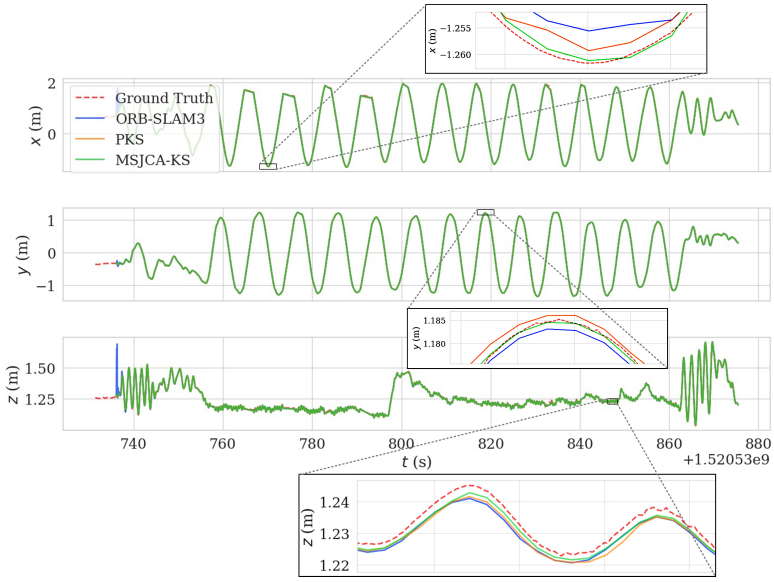

Fig. 9 Comparison of time-varying trajectories in Monocular-Inertial mode for Room2 in the TUM_VI dataset.

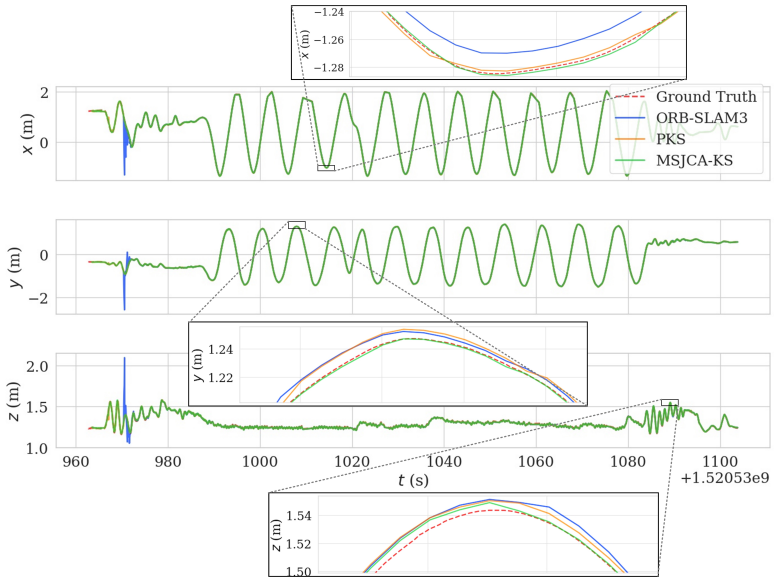

Fig. 10 Comparison of time-varying trajectories in Monocular-Inertial mode for Room3 in the TUM_VI dataset.

4.2 Keyframe comparison

To evaluate the performance of the keyframe selection method proposed in this paper, the MH04 dataset is used to compare with ORB-SLAM3 and PKS algorithms in terms of keyframe quality and quantity. The distribution of selected keyframes in the trajectories of the three algorithms is depicted in Fig. 11 and Fig. 12, with (a) ORB-SLAM3, (b) PKS, and (c) the proposed MSJCA-KS. Notably, Fig. 11 and Fig. 12 reveal that the keyframes of MSJCA-KS exhibit a more uniform distribution compared to the other two algorithms. The disadvantage of the ORB-SLAM3 and PKS algorithms is that they failed to capture some keyframes during fast linear motion, resulting in excessively large intervals between adjacent keyframes. Furthermore, especially during corner motion, both the ORB-SLAM3 and PKS algorithms exhibit keyframe oversampling. The proposed MSJCA-KS ensures better coverage and representation of keyframes, enhancing the overall performance and accuracy of the trajectory estimation.

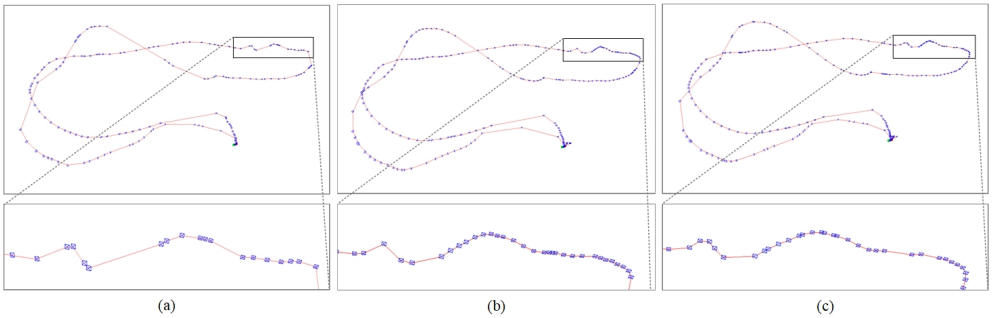

Fig. 11 Comparison of keyframe selection distribution in Stereo-Inertial mode in the EuRoC dataset. (a) ORB-SLAM3, (b) PKS,and (c) the proposed MSJCA-KS

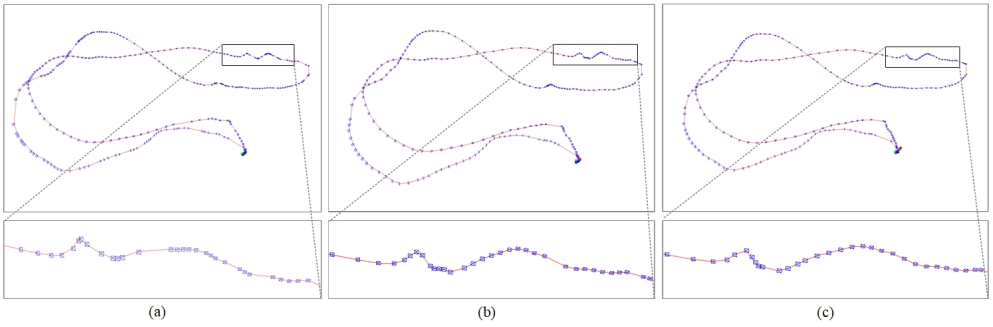

Fig. 12 Comparison of keyframe selection distribution in Monocular-Inertial mode in the EuRoC dataset. (a) ORB-SLAM3, (b) PKS,and (c) the proposed MSJCA-KS

The comparison of keyframe selection among the three methods is presented as followed, with the number of keyframes selected by each algorithm recorded for different

scenarios. To minimize the impact of random errors, each experiment was repeated 10 times, and the average was calculated. Fig. 13 illustrates the number of keyframes for five scenes. In the monocular inertial mode, the proposed MSJCA-KS algorithm consistently selects fewer keyframes compared to the ORB-SLAM3 and PKS algorithms. This outcome strongly demonstrates the superiority of the MSJCA-KS algorithm in terms of keyframe selection accuracy, as well as positioning accuracy and stability. In the stereo-inertial mode, both the PKS algorithm and the proposed MSJCA-KS algorithm select more keyframes than ORB-SLAM3. This can be attributed to the fact that the stereo mode provides stereoscopic vision information, facilitating depth estimation through camera disparities. Consequently, the ORB-SLAM3 system can more easily estimate camera motion, reducing the necessity for keyframes. In the stereo-inertial mode, for both the PKS algorithm and the MSJCA-KS method, keyframe selection primarily relies on photogrammetry and camera geometric constraints, thereby mitigating the influence of depth information on keyframe selection to some extent. It is important to note that, compared to the PKS algorithm, the proposed MSJCA-KS algorithm incorporates a real-time motion monitoring mechanism and utilizes the IMU to optimize and adjust the adaptive threshold. This effectively enhances the algorithm's performance in challenging motion scenarios. And the method proposed in this paper has lower number of selected keyframes compared to the PKS algorithm, while achieving higher accuracy.

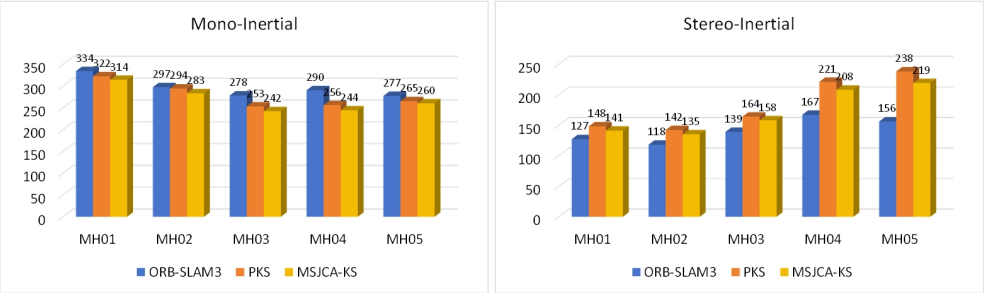

Fig. 13 Comparison of the number of selected keyframes in the EuRoC dataset.

4.3 Running time comparison

The real-time computing capabilities is crucial for visual SLAM systems. To assess the algorithm's performance, we tested the MH01 and MH04 in the EuRoC dataset, and executed each method 10 times in different modes and recorded the average time. The evaluation maintained a time measurement accuracy of 0.1 seconds. Fig. 14 illustrates that both PKS and the proposed method involve the intricate process of establishing a spatial cone for map points, thereby increasing the complexity of keyframe selection compared to the ORB-SLAM3 algorithm, which utilizes a fixed threshold. Although the PKS and proposed MSJCA-KS methods exhibit greater complexity in keyframe selection, the proposed MSJCA-KS method processes fewer keyframes, resulting in a slightly shorter runtime than the PKS algorithm. Nevertheless, these additional computations have a negligible impact on real-time execution, allowing the algorithm to meet the required real-time constraints.

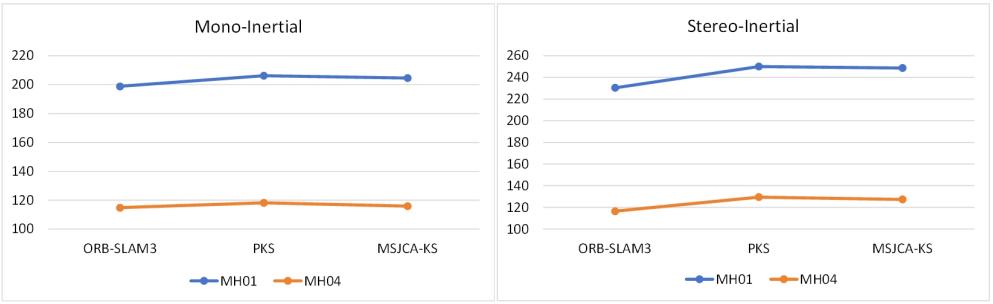

Fig. 14 Comparison of running time(s).

4.4 Ablation experiment

To evaluate the impact of the proposed MSJCA-KS method, we conducted a series of ablation experiments. The RMSE AT is used as a measure of positioning accuracy. The results show that when only camera geometry constraint is used, or only camera geometry and UD constraints are used, the positioning accuracy of the MSJCA-KS method is essentially the same as that of the ORB-SLAM3 or PKS algorithms. However, when the IMU joint constraints are used into the MSJCA-KS method, the positioning accuracy is significantly improved, surpassing the performance of the ORB-SLAM3 or PKS algorithms. These results provide compelling evidence for the effectiveness of IMU joint constraints in improving the accuracy of this method. The experimental results in Table 3, Fig. 15 and Fig. 16 show that the RMSE value gradually decreases as the number of constraints increases, and the estimated trajectory is closer to the ground truth trajectory. This trend indicates the system stability is improved and the positioning accuracy becomes high. Therefore, it can be inferred that all three constraints are essential, as the absence of any one of them could adversely affect the accuracy and stability of the system.

Table 3 RMSE AT results for different combinations(m).

	MH05		MH02	
Sequence	Mono-Inertial	Stereo-Inertial	Mono-Inertial	Stereo-Inertial
Camera+ IMU +UD	0.046	0.040	0.018	0.026
Camera+ IMU	0.047	0.041	0.020	0.026
Camera+ UD	0.052	0.044	0.031	0.033
Camera	0.055	0.045	0.033	0.034

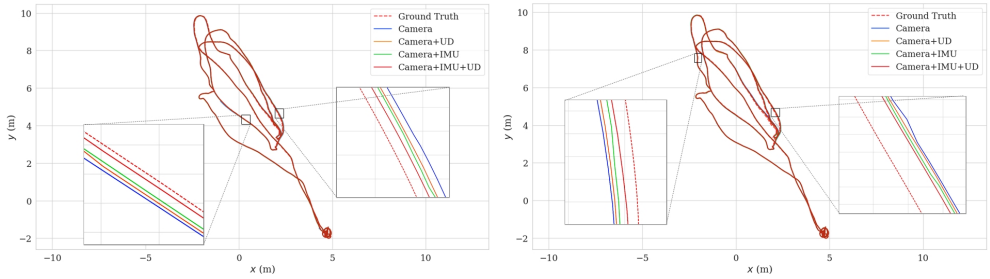

Fig. 15 Comparison of trajectories in different combinations under the MH02 sequence in the EuRoC dataset (monocular mode on the left, stereo mode on the right).

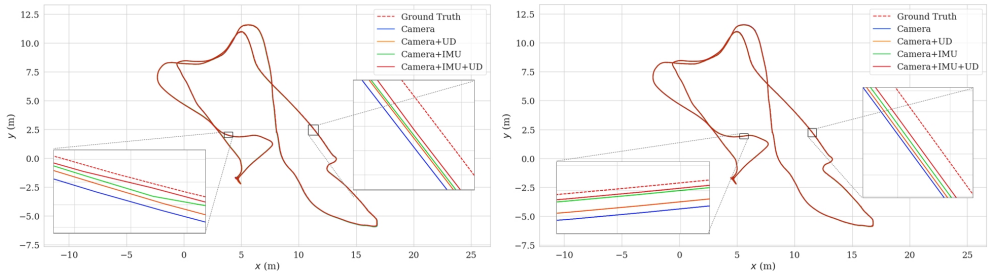

Fig. 16 Comparison of trajectories in different combinations under the MH05 sequence in the EuRoC dataset (monocular mode on the left, stereo mode on the right).

5 Conclusions

This paper proposed a new keyframe selection method for visual SLAM algorithms. It consists of three constraints, including a camera geometric constraint based on the definition of a four-region spatial cone across multiple consecutive frames to improve the adaptability to varying scenes, an IMU joint constraint to improve the robustness when facing rapid camera motions, and an uniform distribution constraint to ensure high quality keyframe selection to improve the stability of SLAM algorithms. The experiments and analyses were conducted by using various sequences from the EuRoC dataset and TUM dataset in both monocular and stereo inertial modes. The results demonstrate that the proposed method significantly improve the positioning accuracy of the visual SLAM system by ensuring adaptability and stability in keyframe selection, particularly in complex environments. Although the real-time performance

of the method slightly lags behind that of ORB-SLAM3 and the influence of depth information on keyframe selection is somewhat diminished, future research can focus on reducing computational complexity to bolster real-time capabilities, and integrating depth information into stereo inertial mode to further improve the positioning accuracy.

References

[1] Sharafutdinov, D., Griguletskii, M., Kopanev, P., Kurenkov, M., Ferrer, G., Burkov, A., Gonnochenko, A., Tsetserukou, D.: Comparison of modern open-source visual slam approaches. Journal of Intelligent & Robotic Systems **107**(3), 43 (2023)

[2] Ebadi, K., Bernreiter, L., Biggie, H., Catt, G., Chang, Y., Chatterjee, A., Denniston, C.E., Deschênes, S.-P., Harlow, K., Khattak, S., et al.: Present and future of slam in extreme environments: The darpa subt challenge. IEEE Transactions on Robotics (2023)

[3] Rebecq, H., Horstschaefer, T., Scaramuzza, D.: Real-time visual-inertial odometry for event cameras using keyframe-based nonlinear optimization (2017)

[4] Wen, W.W., Hsu, L.-T.: 3d lidar aided gnss nlos mitigation in urban canyons. IEEE Transactions on Intelligent Transportation Systems **23**(10), 18224–18236 (2022)

[5] Pan, Y., Zhong, X., Wiesmann, L., Posewsky, T., Behley, J., Stachniss, C.: Pin-slam: Lidar slam using a point-based implicit neural representation for achieving global map consistency. arXiv preprint arXiv:2401.09101 (2024)

[6] Zhao, Y., Vela, P.A.: Good feature matching: Toward accurate, robust vo/vslam with low latency. IEEE Transactions on Robotics **36**(3), 657–675 (2020)

[7] Wen, S., Liu, X., Wang, Z., Zhang, H., Zhang, Z., Tian, W.: An improved multi-object classification algorithm for visual slam under dynamic environment. Intelligent Service Robotics **15**(1), 39–55 (2022)

[8] Wang, S., Zhang, A., Zhang, Z., Zhao, X.: Real-time monocular visual–inertial slam with structural constraints of line and point–line fusion. Intelligent Service Robotics, 1–20 (2023)

[9] Azimi, A., Ahmadabadian, A.H., Remondino, F.: Pks: A photogrammetric key-frame selection method for visual-inertial systems built on orb-slam3. ISPRS Journal of Photogrammetry and Remote Sensing **191**, 18–32 (2022)

[10] Sheng, L., Xu, D., Ouyang, W., Wang, X.: Unsupervised collaborative learning of keyframe detection and visual odometry towards monocular deep slam. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp.

4302–4311 (2019)

[11] Lu, P.J., Chuang, J.-H.: Fusion of multi-intensity image for deep learning-based human and face detection. IEEE Access **10**, 8816–8823 (2022)

[12] Klein, G., Murray, D.: Parallel tracking and mapping for small ar workspaces. In: 2007 6th IEEE and ACM International Symposium on Mixed and Augmented Reality, pp. 225–234 (2007). IEEE

[13] Forster, C., Pizzoli, M., Scaramuzza, D.: Svo: Fast semi-direct monocular visual odometry. In: 2014 IEEE International Conference on Robotics and Automation (ICRA), pp. 15–22 (2014). IEEE

[14] Engel, J., Schöps, T., Cremers, D.: Lsd-slam: Large-scale direct monocular slam. In: European Conference on Computer Vision, pp. 834–849 (2014). Springer

[15] Endres, F., Hess, J., Sturm, J., Cremers, D., Burgard, W.: 3-d mapping with an rgb-d camera. IEEE transactions on robotics **30**(1), 177–187 (2013)

[16] Lee, G.H., Fraundorfer, F., Pollefeys, M.: Rs-slam: Ransac sampling for visual fastslam. In: 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 1655–1660 (2011). IEEE

[17] Thormählen, T., Broszio, H., Weissenfeld, A.: Keyframe selection for camera motion and structure estimation from multiple views. In: Computer Vision-ECCV 2004: 8th European Conference on Computer Vision, Prague, Czech Republic, May 11-14, 2004. Proceedings, Part I 8, pp. 523–535 (2004). Springer

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

[19] Tan, W., Liu, H., Dong, Z., Zhang, G., Bao, H.: Robust monocular slam in dynamic environments. In: 2013 IEEE International Symposium on Mixed and Augmented Reality (ISMAR), pp. 209–218 (2013). IEEE

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

[21] Kerl, C., Sturm, J., Cremers, D.: Dense visual slam for rgb-d cameras. In: 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 2100–2106 (2013). IEEE

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

[23] Fanfani, M., Bellavia, F., Colombo, C.: Accurate keyframe selection and keypoint

tracking for robust visual odometry. Machine Vision and Applications **27**, 833–844 (2016)

[24] Stalbaum, J., Song, J.-b.: Keyframe and inlier selection for visual slam. In: 2013 10th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI), pp. 391–396 (2013). IEEE

[25] Engel, J., Koltun, V., Cremers, D.: Direct sparse odometry. IEEE transactions on pattern analysis and machine intelligence **40**(3), 611–625 (2017)

[26] Gao, X., Wang, R., Demmel, N., Cremers, D.: Ldso: Direct sparse odometry with loop closure. In: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 2198–2204 (2018). IEEE

[27] Zubizarreta, J., Aguinaga, I., Montiel, J.M.M.: Direct sparse mapping. IEEE Transactions on Robotics **36**(4), 1363–1370 (2020)

[28] Dong, Z., Zhang, G., Jia, J., Bao, H.: Keyframe-based real-time camera tracking. In: 2009 IEEE 12th International Conference on Computer Vision, pp. 1538–1545 (2009). IEEE

[29] Chen, W., Zhu, L., Lin, X., He, L., Guan, Y., Zhang, H.: Dynamic strategy of keyframe selection with pd controller for vslam systems. IEEE/ASME Transactions on Mechatronics **27**(1), 115–125 (2021)

[30] Wang, J., Zeng, C., Wang, Z., Jiang, K.: An improved smart key frame extraction algorithm for vehicle target recognition. Computers & Electrical Engineering **97**, 107540 (2022)

[31] Campos, C., Elvira, R., Rodríguez, J.J.G., Montiel, J.M., Tardós, J.D.: Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam. IEEE Transactions on Robotics **37**(6), 1874–1890 (2021)

[32] Khattak, S., Papachristos, C., Alexis, K.: Keyframe-based thermal–inertial odometry. Journal of Field Robotics **37**(4), 552–579 (2020)

[33] Sousa, R.B., Sobreira, H.M., Moreira, A.P.: A systematic literature review on long-term localization and mapping for mobile robots. Journal of Field Robotics **40**(5), 1245–1322 (2023)

[34] Hosseininaveh, A., Serpico, M., Robson, S., Hess, M., Boehm, J., Pridden, I., Amati, G.: Automatic image selection in photogrammetric multi-view stereo methods. (2012). Eurographics Association

[35] Azimi, A., Hosseininaveh, A., Remondino, F.: a novel geometric key-frame selection method for visual-inertial slam and odometry systems. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information

Sciences **43**, 9–14 (2022)

[36] Gomez-Ojeda, R., Moreno, F.-A., Zuniga-Noël, D., Scaramuzza, D., Gonzalez-Jimenez, J.: Pl-slam: A stereo slam system through the combination of points and line segments. IEEE Transactions on Robotics **35**(3), 734–746 (2019)