Chinese Journal of Mechanical Engineering

ORIGINAL ARTICLE

Open Access

Human Visual Attention Mechanism-Inspired Point-and-Line Stereo Visual Odometry for Environments with Uneven Distributed Features

Chang Wang[1], Jianhua Zhang[1], Yan Zhao[1*], Youjie Zhou[2] and Jincheng Jiang[3]

Abstract

Visual odometry is critical in visual simultaneous localization and mapping for robot navigation. However, the pose estimation performance of most current visual odometry algorithms degrades in scenes with unevenly distributed features because dense features occupy excessive weight. Herein, a new human visual attention mechanism for point-and-line stereo visual odometry, which is called point-line-weight-mechanism visual odometry (PLWM-VO), is proposed to describe scene features in a global and balanced manner. A weight-adaptive model based on region partition and region growth is generated for the human visual attention mechanism, where sufficient attention is assigned to position-distinctive objects (sparse features in the environment). Furthermore, the sum of absolute differences algorithm is used to improve the accuracy of initialization for line features. Compared with the state-of-the-art method (ORB-VO), PLWM-VO show a 36.79% reduction in the absolute trajectory error on the Kitti and Euroc datasets. Although the time consumption of PLWM-VO is higher than that of ORB-VO, online test results indicate that PLWM-VO satisfies the real-time demand. The proposed algorithm not only significantly promotes the environmental adaptability of visual odometry, but also quantitatively demonstrates the superiority of the human visual attention mechanism.

Keywords Visual odometry, Human visual attention mechanism, Environmental adaptability, Uneven distributed features

1 Introduction

Research pertaining to visual odometry has primarily focused on improving its position estimation accuracy in various environments [1–4]. To improve the adaptability of visual odometry to low-texture and point-rich environments, several algorithms have been developed to describe the environment based on different types of features, including points [5], lines [6, 7], planes [8], edges [9], and cluster feature [10]. However, current methods typically do not offer sufficient adaptability to environments with unevenly distributed features, which are ubiquitous in the real world. This may result in the failure of robot placement and navigation.

For point-based methods, the extended Kalman filter was used to approximate point feature information to a linear Gaussian to achieve simultaneous placement and map construction in Mono simultaneous localization and mapping (SLAM) [11]. In Klein's parallel tracking and mapping algorithm [12], tracking and mapping are categorized into two parallel threads to realize vision localization. A depth camera was used to obtain point features

*Correspondence:
Yan Zhao
zhaoyan-0312@hotmail.com
[1] School of Mechanical Engineering, University of Science and Technology Beijing, Beijing 100083, China
[2] School of Mechanical Engineering, Shandong University, Jinan 250100, China
[3] School of Mechanical Engineering, Hebei University of Technology, Tianjin 300401, China

with depth information in RGB-D SLAM [13]. The iterative closest point (ICP) algorithm was used to optimize the point cloud to obtain an accurate pose estimation. In 2015, Mur-Artal et al. [14] proposed an ORB-SLAM algorithm, which is based on oriented FAST and rotated BRIEF (ORB) [15]. The pose was estimated by minimizing the reprojection error of the points using the least-squares method. However, these methods, which are based only on point features, cannot fully describe low-texture environments. Therefore, point-based methods are not suitable for low-texture environments where point features are sparse. Point-line fusion methods have been proposed to improve the adaptability of visual odometry in both point-rich and low-texture environments. Witt et al. [16] proposed an iterative closest multiple line algorithm, and the ICP algorithm was used for the line segments. This improves the pose estimation accuracy, even at a small rotation angle. Lu et al. [17] proposed a visual odometry scheme that combines points and line features; it detects line segments in the RGB-D camera images. Li et al. [18] proposed a new algorithm to integrate key points into line segments. Key points were selected by referring to the line segments. To achieve better pose estimation performance, Gomez-Ojeda et al. [19, 20] proposed a point-and-line-based stereo visual odometry (PL-SVO) algorithm, which assigns weights between point and line features based on the number of points and lines. Meanwhile, the stability of the line-segment detector (LSD) algorithm [21] was improved by distinguishing adjacent line segments and merging them [22]. However, because the globality and balance of feature descriptions are disregarded, both methods based on point features and point-line features cannot adapt well to scenes with unevenly distributed features. This reduces the location accuracy and robustness of visual odometry in scenes that are ubiquitous in real-world environments.

Feature-sparse regions may contain more robust features in real-world scenes, whereas feature-dense regions may include many invalid features because of feature redundancy. In the pose estimation of current visual odometry algorithms, many feature-robust regions do not receive sufficient priority, whereas many feature-invalid regions have garnered significant attention. Such an unreasonable weight assignment method affects the globality and balance of the feature descriptions, thereby reducing the accuracy of pose estimation, particularly in environments with unevenly distributed features.

Although ORB-SLAM2 [23], which filters point features using a quadtree, has been proposed, it does not solve the problem of uneven distributed feature descriptions adequately. It can realize a uniform distribution of feature points to a certain extent. However, it is easy to produce weak point features by just changing the detecting threshold in local regions. This results in more mismatched pairs when using these weak point features, which reduces the accuracy of the pose estimation of visual odometry. In addition, the algorithm is only suitable for point features. This is because the selection method for local features in the quadtree cannot be used for nonlocal features and line segments. This limits the capability of ORB-SLAM2 in low-texture environments, where line features are abundant, whereas point features are scarce.

In the human visual attention mechanism, distinctive objects that exhibit different characteristics are prioritized over the surrounding attributes, such as different positions, colors, and shapes [24]. The visual attention mechanism enables humans to process information in feature-dense and -sparse regions in a global and balanced manner [25, 26]. Therefore, humans can achieve good performances in localization, target recognition, and scene understanding, including in environments with unevenly distributed features. It provides a promising way to further improve the environmental adaptability of the visual odometry.

Inspired by the human visual attention mechanism, we herein propose stereo visual odometry, abbreviated as point-line-weight-mechanism visual odometry (PLWM-VO). Our main contributions are as follows:

(1) Human visual attention mechanism that prioritize distinctive objects is introduced into visual odometry for improving its adaptability to environments with uneven distributed features. A weight adaptive model is proposed to achieve good globality and balance between feature-dense and -sparse regions.
(2) A modified reprojection error model is developed using the sum of absolute differences (SAD) [27] algorithm to further improve the accuracy of pose estimation in visual odometry.
(3) The human visual attention mechanism for improving the adaptability of visual odometry is verified in both low-texture and point-rich environments through experiments on public datasets and online tests. The results indicate that the proposed PLWM-VO achieves better performance than the state-of-the-art method ORB-VO (ORB-SLAM2 without loop closing).

The remainder of this paper is organized as follows: The proposed PLWM-VO algorithm is described comprehensively in Section 2. Section 3 describes the experiments conducted and presents the results obtained. The performance of the proposed algorithm is further discussed in Section 4. Finally, conclusions and future studies are presented in Section 5.

2 Proposed PLWM-VO Algorithm

Uneven distributions of features are ubiquitous in real worlds, where most existing point-line fusion-based visual odometry methods demonstrate low adaptability. They assign the same weight to all features in both feature-sparse and -dense regions. Owing to the mass of features in feature-dense regions, these areas typically feature an excessive weight. Consequently, the feature-sparse regions are assigned a small weight. In other words, the current methods cannot balance the weight assignment between feature-dense and -sparse regions in the scene. In addition, the features of the entire scene are not used reasonably for pose estimation. Consequently, current visual odometry methods cannot accurately perform pose estimation in environments with unevenly distributed features. However, the human visual attention mechanism exhibits good environmental adaptability for visual information processing. It prioritizes distinctive objects in the scenes based on their position, color, shape, etc. The feature positions in feature-spare regions are relatively distinctive compared with those in feature-dense regions. Thus, the features in feature-sparse regions are typically prioritized. Such a visual attention mechanism allows humans to use both feature-dense and -sparse regions in the scene in a global and balanced manner. To improve the adaptability of visual odometry for scenes with unevenly distributed features, from the perspective of the position distinction of features, we introduce the human visual attention mechanism into visual odometry. We establish a PLWM-VO algorithm with improved globality and balance of feature description. Figure 1 shows the overall framework of the PLWM-VO algorithm.

(1) First, point features and line features are selected and associated with the ORB and LSD [21]-LBD [28] algorithms in parallel, respectively. Mismatches are eliminated using the random sample consensus RANSAC [29, 30] algorithm.
(2) Second, a weight-adaptive model is proposed to introduce the human visual attention mechanism into the cost function, which is used to calculate the pose increment between adjacent frames through iterative minimization. The improvement in the

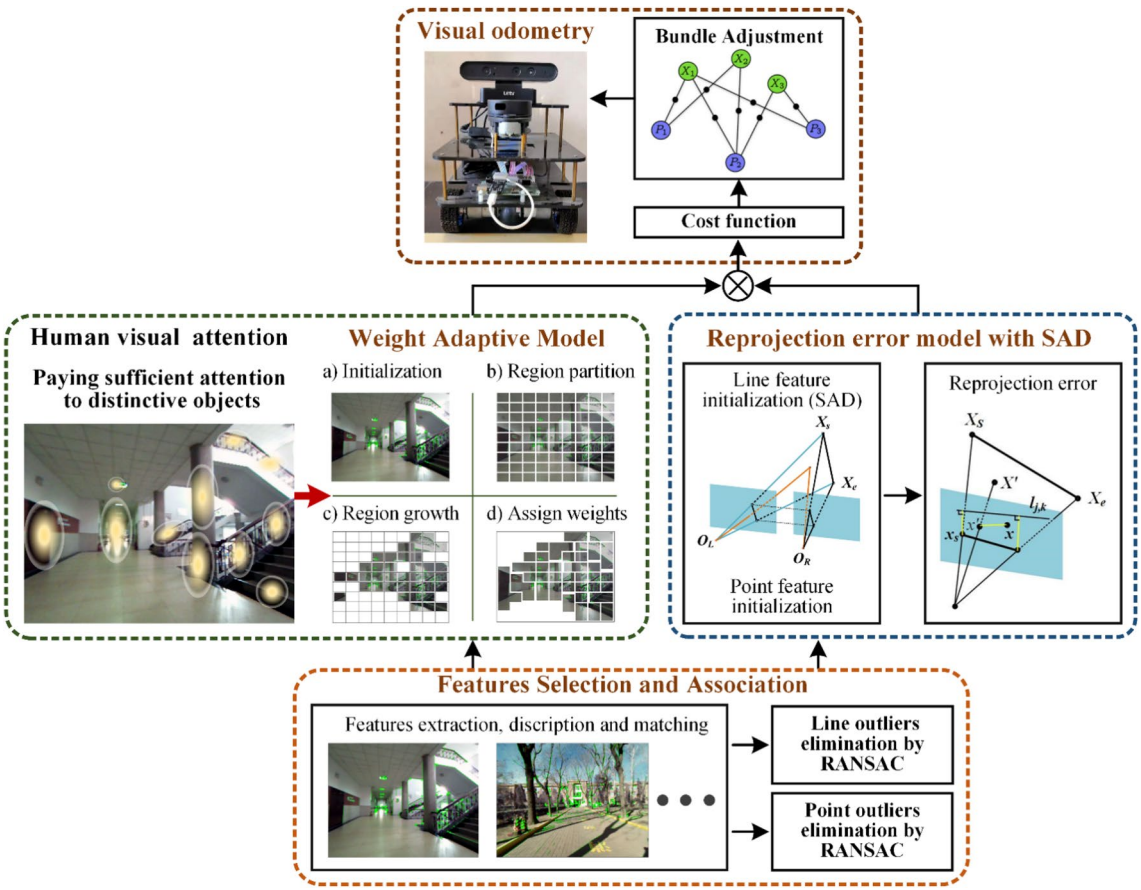

Figure 1 Overview of proposed PLWM-VO algorithm

globality and balance of the feature description is primarily realized using the weight-adaptive model.

(3) Third, a modified reprojection error model is proposed. The accuracy of initialization is enhanced by accurately detecting the line endpoints using the SAD algorithm.

(4) Fourth, the final pose estimation is realized by calculating the cumulative sum of pose increments between adjacent frames using the Gauss–Newton method.

2.1 Pose Estimation via Iterative Minimization of Cost Function

The pose increment ξ between two adjacent frames is calculated by the iterative minimization of the cost function, as expressed by in Eq. (1). During this process, the Gauss–Newton method is used to solve the incremental equation linearly, as shown in Eq. (2). The probability of observing data becomes the maximum. In this process, if iteration increment Δx_k is sufficiently small, then the calculation is terminated. Otherwise, we set $\xi_{k+1} = \xi_k + \Delta x_k$ and continue to optimize iteratively.

$$\xi^* = \arg\min E(\xi), \tag{1}$$

$$\boldsymbol{H}\Delta X = g, \tag{2}$$

$$\boldsymbol{H} = \boldsymbol{J}_{ij}^{\mathrm{T}} \boldsymbol{J}_{ij}, \tag{3}$$

$$g = -\boldsymbol{J}_{ij} f(x), \tag{4}$$

$$E(\xi) = ||f(x)||^2, \tag{5}$$

where $\boldsymbol{J}_{ij}$ represents the Jacobian matrix corresponding to the cost function $E(\xi)$. $\boldsymbol{H}$ is used as an approximation of the second-order Hessian matrix in the Gauss–Newton method, as shown in Eq. (3). $E(\xi)$ is the absolute value of $f(x)$ squared.

The cost function $E(\xi)$ is composed of the reprojection error of the point features and the endpoints of the line features, as shown in Eq. (6). $\sum_{e_{i,k}^p}^{-1}$ and $\sum_{e_{j,k}^l}^{-1}$ represent the inverse of the covariance matrix corresponding to the reprojection error of the ith point feature and jth line feature, respectively. $\boldsymbol{H}_p$ and $\boldsymbol{H}_l$ are the Huber robust kernel functions corresponding to the points and lines, respectively. Here, ω represents the weight corresponding to the reprojection error of the points and lines. A reasonable distribution of weights is important for pose estimation.

$$\begin{aligned} E(\xi) = &\sum_{i=0}^{\mathbf{N}_p} \omega_p (\boldsymbol{H}_p\, e_{i,k}^p \sum\nolimits_{e_{i,k}^p}^{-1} e_{i,k}^p) \\ &+ \sum_{j=0}^{\mathbf{N}_l} \omega_{xs} (\boldsymbol{H}_l\, e_{j,k}^{l_{xs}} \sum\nolimits_{e_{j,k}^l}^{-1} e_{i,k}^{l_{xs}}) \\ &+ \sum_{j=0}^{\mathbf{N}_l} \omega_{xe} (\boldsymbol{H}_l\, e_{j,k}^{l_{xe}} \sum\nolimits_{e_{j,k}^l}^{-1} e_{i,k}^{l_{xe}}). \end{aligned} \tag{6}$$

2.2 Proposed Weight-Adaptive Model

When visual odometry is performed in actual application scenes, the features are typically dense in some regions but sparse or absent in others. The current visual odometry method assigns the same weight to the extracted features. Consequently, the weight of the feature-dense regions becomes excessive, whereas the weight of the feature-sparse regions is insignificant. The region partition method was adopted to solve this problem. Images were partitioned into $n \times n$ grids of the same size. The weights of the feature-dense and -sparse regions no longer depend on the number of features in the region but rather on the size of the region. Therefore, the feature-dense and -sparse regions have the same weight.

In the human visual attention mechanism, more weight is typically assigned to distinctive objects. In visual odometry, the position of features in feature-sparse regions is distinctive compared with that of features in feature-dense regions. Furthermore, feature-sparse regions contain only a few features distributed sparsely in a certain area. Within a certain size range, the smaller the number of features, the greater is the degree of sparsity of the area. Meanwhile, the greater the distinction of the features in the feature-sparse regions, the greater is the weight of the feature-sparse regions. Therefore, the region-growth method was used in this study. Grids without features were merged into effective grids that contained features. The weight coefficients of the effective grids in the sparse region were improved adaptively. Hence, the globality and balance of the feature descriptions were realized based on the feature positions. Finally, a weight-adaptive model based on region partition and region growth was established to adapt to a scene with unevenly distributed features. The specific process of the weight-adaptive model is as follows:

(1) *Image initialization*: The total number of features n_f (including points and lines) and the corresponding

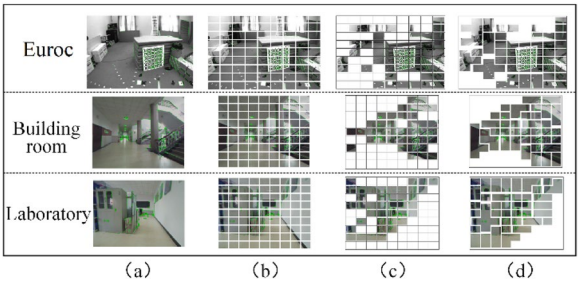

Figure 2 Four stages of weight-adaptive model: **a** Image initialization, **b** region partition, **c** region growth, **d** weight allocation based on region

pixel coordinates are calculated, as shown in Figure 2a.

(2) *Region partition*: The image is partitioned into $n \times n$ small grids. The features are stored in each grid unit, S_i. The number of endpoints S_i is calculated as shown in Figure 2b.

(3) *Region growth*: The region growth method is used to merge the grids, as shown in Figure 2c and d. Grids without features are defined as near (S). Grids that contain features are denoted as S_m, S_{m+1}…. The objects for region growth are near (S) and adjacent to S_m, S_{m+1}…. During region growth, the following four principles must be complied:

- *Principle 1*: When the region near (S_i) has no adjacent seed grids, the grid near (S_i) is considered invalid, as shown in Figure 3a.
- *Principle 2*: When the region near (S) has adjacent seed grids S_m, S_{m+1}…, then the number of seed grids is larger than or equal to two. If the number of features (S_m, S_{m+1}…) in each seed grid is different, then the seed grid with the most features is grown, as shown in Figure 3b.
- *Principle 3*: When the region near (S) has an adjacent seed grid S_m and the number of seed grids is one, then the region is enlarged, as shown in Figure 3c.

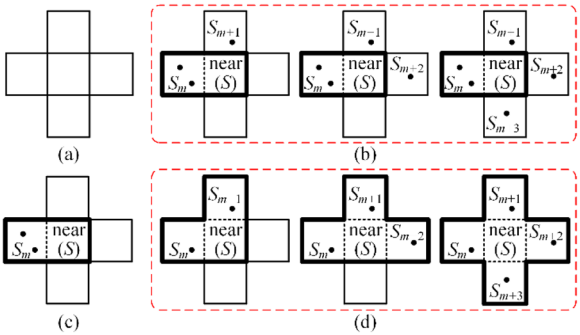

Figure 3 Four stages of weight-adaptive model: **a** Case 1, **b** Case 2, **c** Case 3, and **d** Case 4

- *Principle 4*: When the region near (S) has adjacent seed grids, then the number of seed grids is larger than or equal to two. If the number of features in each seed grid is the same, then the seed grids and the grids near (S) are merged, as shown in Figure 3d.

(4) *Weight allocation*: The features are weighted based on the proportion of the corresponding grid area after the region growth, as shown in Table 1. If the proportion of a certain grid in the entire picture is q and the number of features in the grid (including the point features and the endpoints of the line features) is S_m, then the weight ω corresponding to each feature in the grid is expressed as follows:

$$\omega = q \cdot (1/S_m). \tag{7}$$

2.3 Modified Reprojection Error Model Using SAD

2.3.1 Improved Feature Initialization Using SAD

Feature initialization is a prerequisite for the reprojection error model prepared for pose estimation. Features must be initialized in a three-dimensional space. Subsequently, the reprojection error is calculated for pose estimation. The triangulation algorithm can be used for initializing point features, but not for initializing line features.

For line features, which are generated by the different illumination conditions in the left and right cameras, the left and right images deviate on the pixels. However, the line feature regions detected by the LSD algorithm cannot not be accurately matched in the left and right images. The results of direct initialization of the detected line-segment endpoints are different from the actual values. This problem is currently solved using the alignment method, which is used to determine the approximate corresponding endpoints to reduce the

Table 1 Algorithm for weight-adaptive model

Input	Pixel coordinates of x_p, x_s, and x_e; feature number n_f
Output	Adaptive coefficients of weights ω
1:	for $i \leftarrow 1$ to n_f do
2:	$C_j \leftarrow$ The pixel coordinates(ORB,LSD)
3:	if ($C_j \in S_i$) then s_i++, $S_i \leftarrow C_j$
4:	end
5:	for $i \leftarrow 1$ to n^2 do
6:	if $s_i \neq 0$ then $S_k \leftarrow S_i$, k++
7:	end
8:	for $m \leftarrow 1$ to k do
9:	if $(s_{i\cdot near(S)} = 0)\&\&(s_m > s_{i\cdot near(near(S))})$
10:	then $S_m \leftarrow S_m +$ near(S)
11:	end
12:	$\omega = q \cdot (1/s_m)$
13:	return ω

initialization error. However, because of the different viewpoints on the left and right, the shapes and positions of the same space line are different in the left and right images. Consequently, the endpoints obtained using the alignment method are different from the actual endpoints. The deviation increases with the angle between the line segment and vertical direction. Therefore, the result of this initialization method is different from the actual value, as shown in Figure 4a.

For improving the accuracy of the initialization of line features, ensuring the adaptability of line features initialization in different environments, it is considered that such a hypothesis is reasonable: if the gray-level distribution of a certain local region within an area to be tested on the right is the closest to that of the line feature endpoint on the left, then it can be regarded as the most accurate corresponding endpoint. Such a local region can be detected using the SAD algorithm [27], which is expressed in Eq. (8). Figure 4b shows that the detected point differs significantly from the actual point when alignment constraints are used. The endpoints can be accurately detected using the SAD:

$$\mathrm{SAD}(u,v)=\sum_{u=1}^{5}\sum_{v=1}^{5}\left|left(u,v)-right(u,v)\right|. \tag{8}$$

The detailed procedure is as follows:

First, the alignment constraint is used to determine the two corresponding approximate endpoints after the matching process. Second, the center of the right endpoint is defined as the center of the detected block (5 × 21 pixels). Finally, a sliding window of 5 × 5 pixels is used to scan the detected block to determine the target local region that is the most similar to the left endpoint, as shown in Figure 4c. Figure 4d shows the pixel difference curve corresponding to Figure 4c.

2.3.2 Reprojection Error Model

The reprojection error of the point features can be obtained by calculating the error between the projection and detected points. As shown in Eq. (9), $x_{i,k}$ and $x'_{i,k}$ represent the ith detected point and projected point in Frame$_k$, respectively.

$$e^{p}_{i,k}=x_{i,k}-x'_{i,k}. \tag{9}$$

For line features, the reprojection errors are defined by measuring the distance between the two endpoints of the projected line segments to the corresponding segments detected. As shown in Eqs. (10) and (11), $l_{j,k}$ represents the jth detected line segment in Frame$_k$. x_s and x_e represent the two endpoints of the projected line segment; $e^{l_{xs}}_{j,k}$ and $e^{l_{xe}}_{j,k}$ represent the corresponding reprojection errors of the two endpoints, respectively.

$$e^{l_{xs}}_{j,k}=\mathrm{d}(x_s,l_{j,k})=[l^{\mathrm{T}}_{j,k}\cdot x_s^{l'_{j,k}}], \tag{10}$$

$$e^{l_{xe}}_{j,k}=\mathrm{d}(x_e,l_{j,k})=[l^{\mathrm{T}}_{j,k}\cdot x_e^{l'_{j,k}}]. \tag{11}$$

Figure 5 shows the process to calculate the reprojection error: Frame$_k$, Frame$_{k-1}$, and Frame$_{k-2}$ represent continuous keyframes, the blue line passing through them represents the estimated trajectory, X_s and X_e represent the projected line segments, and the yellow line segments represent the distance from the endpoints of the projected line segments to the detected lines. This method allows the reprojection error to be described steadily

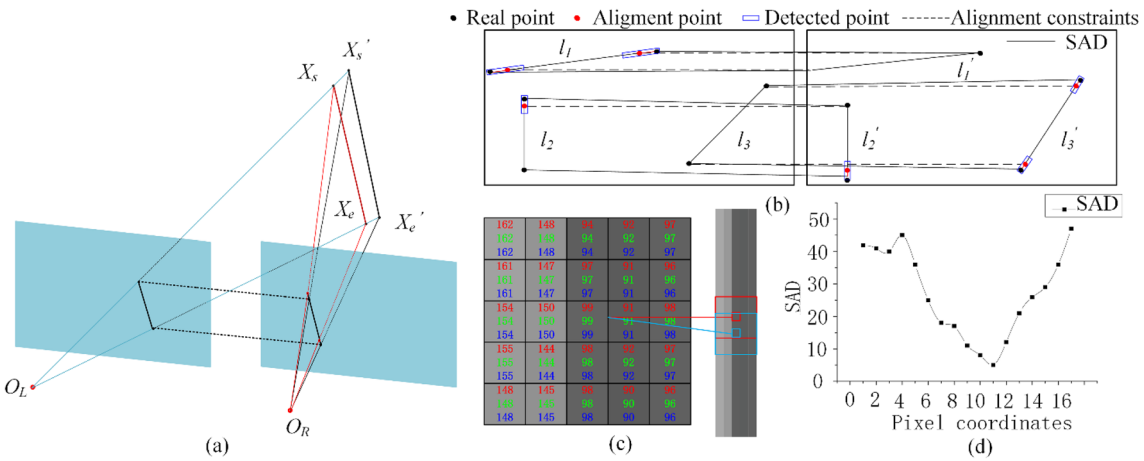

Figure 4 Improved feature initialization using SAD: **a** Initialization of line features, **b** difference between the two methods, **c** registration effect of two methods, **d** corresponding value of SAD for line features

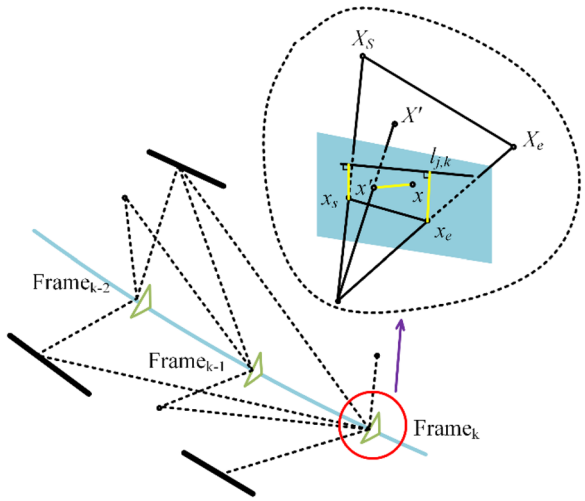

Figure 5 Process for calculating reprojection error in point-line features

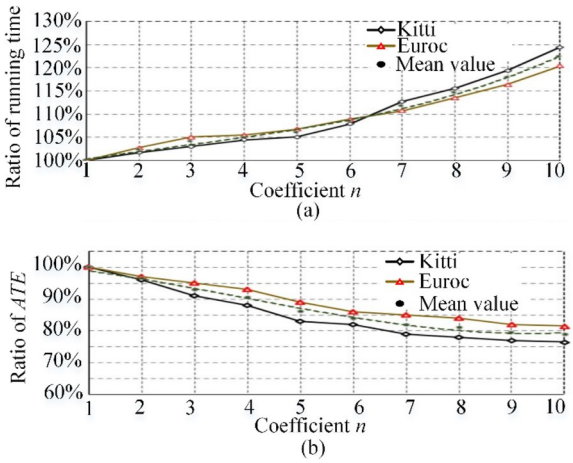

Figure 6 Average timing and ATE results with changes of n: **a** Timing ratios and **b** ATE ratios

because the projected line segments are obtained from space line segments that are initialized by accurately detecting endpoints using the SAD algorithm.

In pose estimation, the change in pose between frames is generally described by the reprojection error model. The cost function is obtained by multiplying the reprojection error model established above with the weight coefficient ω obtained using the weight-adaptive model proposed in Section 2.2. Subsequently, the cost function is iteratively minimized to obtain the pose increment between frames. Finally, pose estimation is achieved by calculating the sum of the pose increments.

3 Experiment and Results

To evaluate the performance of the proposed PLWM-VO algorithm, comparative experiments on open datasets and online tests were conducted. First, the effect of the region partition grid number n on the pose-estimation accuracy was analyzed. Second, comparative tests between the proposed PLWM-VO and state-of-the-art methods, ORB-VO (ORB-SLAM2 without loop closing) and PL-SVO, were conducted on the KITTI [31] and Euroc [32] datasets. The absolute trajectory error (ATE) [33] was used as the evaluation metric, which is a quantitative representation of the difference between the estimated trajectory and ground truth. Third, online tests were conducted in several real-world environments. All experiments were performed on a laptop equipped with an Intel Core i7-7500U CPU at 2.90 GHz, 8 GB of RAM, and Ubuntu 16.04.

3.1 Experiments on Open Datasets

3.1.1 Effect of Region Partition Grid Number n on Algorithm Performance

In this section, we analyze the effect of the region partition grid number n on the operating time and pose estimation accuracy of the proposed PLWM-VO. We set n from 1 to 10. The operating time and ATE were used as the evaluation metrics. For convenience, we denote α_i and β_i as the averages of the ratios of $time_{i,j}$ and $ATE_{i,j}$, respectively, in the m sequences of the Kitti and Euroc datasets. Here, the time ratio is $time_{i,j}$ divided by $time_{1,j}$. Meanwhile, the ATE ratio is $ATE_{i,j}$ divided by $ATE_{1,j}$. α_i and β_i can be expressed by Eqs. (12) and (13), respectively, where $time_{i,j}$ and $ATE_{i,j}$ represent the operating time and ATE, respectively, of the proposed PLWM-VO on the jth dataset when $n = i$.

$$\alpha_i = 1/m \times \sum_{j=1}^{m} \frac{time_{i,j}}{time_{1,j}}, \tag{12}$$

$$\beta_i = 1/m \times \sum_{j=1}^{m} \frac{ATE_{i,j}}{ATE_{1,j}}. \tag{13}$$

α_i and β_i were calculated separately for the Kitti and Euroc datasets. Figure 6 shows the curves of α_i and β_i vs. n. As shown in Figure 6a, as n increases, α_i increases continuously under the Kitti and Euroc datasets, and the growth rate increases gradually. This is because as n increases, the number of grids obtained by dividing each image in the weight adaptive model increases. Consequently, a higher computational cost is incurred during the calculation, which increases the operating time significantly.

As shown in Figure 6b, as n increases, β_i decreases continuously, but the reduction rate decreases. This is because as n increases, the distinctive features are assigned more weights. A global and balanced manner for feature description is gradually realized, and the accuracy of pose estimation is improved. As n increased from 1 to 9, β_i decreased significantly; however, the difference between β_9 and β_{10} was insignificant. This is because as n increases continuously, the feature description can be performed in a global and balanced manner. However, when n increases to a certain degree, with the aid of the weight-adaptive model, the degree of attention to the sparse area of the visual odometry stabilizes, and the rate of increase in the weight of the sparse area decreases. The continued growth does not improve the accuracy of the pose estimation, but increases the operating time. Therefore, n was set to 9 to improve the performance of the proposed algorithm.

As shown in Figure 7, we compared and analyzed the difference in weights of the feature-sparse and -dense regions for $n = 1$ and $n = 9$. The ω(s) values shown in Figure 7d, e, and f are greater than those shown in Figure 7a, b, and c by 15.0%, 2.2%, and 2.5%, respectively. The chairs and windows of the feature-sparse regions in Figure 7 are position specific in the scenes. This shows that the weight of the specific features in the sparse region is improved when $n = 9$ compared with that when $n = 1$. This verifies that the weight-adaptive model is applicable to scenes with unevenly distributed features.

The differences between the weight ω(s) of the feature-sparse regions and the ω(d) of the feature-dense region in Figure 7a, b, and c were 28.1%, 31.6%, and 34.8%, respectively. The corresponding differences in Figure 7d, e, and f were 5.4%, 15.0%, and 11.3%, respectively. The weight gap between the feature-sparse and -dense regions was smaller when $n = 9$ as compared with when $n = 1$. This verifies that the use of the weight-adaptive model improves the globality and balance for scenes with unevenly distributed features.

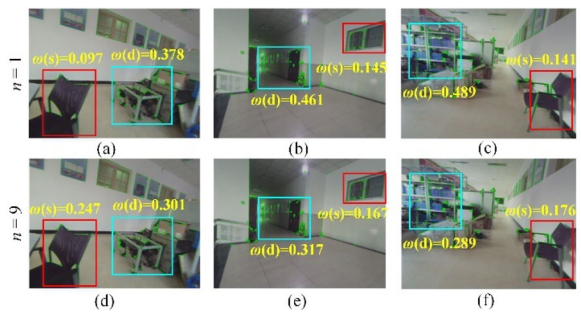

Figure 7 Weights of feature-sparse and -dense regions with n set to 1 and 9: **a** case 1 with n = 1, **b** case 2 with n = 1, **c** case 3 with n = 1, **d** case 1 with n = 9, **e** case 2 with n = 9 and **f** case 3 with n = 9

As shown in Figure 6b, our method reached a general optimal state for pose estimation when $n = 9$. As shown in Figures 7d, e, and f, the weights ω(d) of the feature-dense regions and the ω(s) of the feature-sparse regions are not completely equal. This proves that achieving good balance does not guarantee an absolute average weight in each region. Therefore, the globality and balance of the weight distribution in a scene with unevenly distributed features is a relative concept.

3.1.2 Method Comparison

As described in Sect. 2, line and point features can be detected stably in both low-textured and point-rich environments. In this study we selected a region partition grid number n of 9. Figure 8 shows some frames in the Kitti and Euroc dataset outputs of the proposed PLWM-VO. They indicate that the environment can be described by important geometric information obtained from the points and line features.

The performances of PLWM-VO, PL-SVO, and ORB-VO were compared with those of the Kitti and Euroc datasets. Eight heat maps were generated (as shown in Figure 9) to compare the trajectory generated by the proposed PLWM-VO with the ground truth. In the heat maps, red and blue correspond to higher and lower error levels, respectively. The gray dashed line represents the ground truth, and the other line represents the trajectory yielded by the proposed PLWM-VO. The error is expressed in units of meter. A comparison shows that the results yielded by the proposed algorithm are similar to the ground truth.

The error measures, i.e., the root mean square error (RMSE) of PLWM-VO, PL-SVO, and ORB-VO for the Kitti and Euroc datasets are listed in Table 2. To verify the usefulness of the SAD algorithm for pose estimation, the test results of the PLWM-VO-without-SAD algorithm

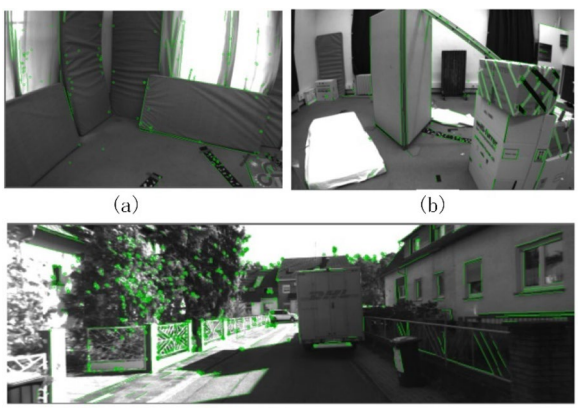

Figure 8 Output of proposed PLWM-VO: **a** case 1, **b** case 2, **c** case 3

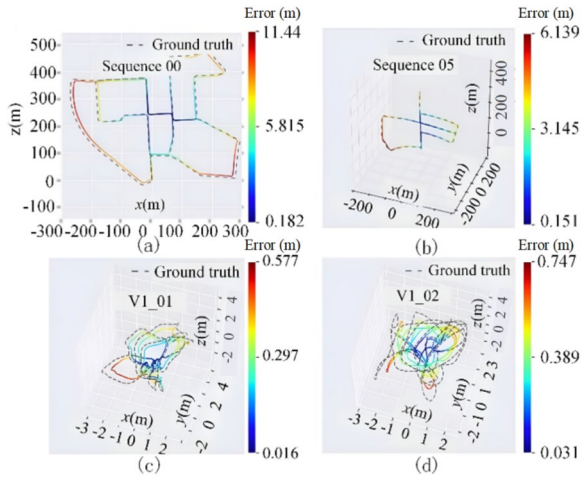

Figure 9 Four heat maps yielded by ORB-VO and the proposed method: **a** results on Kitti00, **b** results on Kitti05, **c** results on V101, **d** results on V102

Table 2 Results of method comparison based on ATE RMSE (m)

Sequences	PL-SVO	ORB-VO	PLWM-VO without SAD	PLWM-VO
Kitti00	13.58	8.35	6.56	**6.47**
Kitti01	150.53	180.47	110.72	**106.56**
Kitti02	58.29	19.73	16.37	**15.60**
Kitti03	13.51	2.66	2.42	**2.38**
Kitti04	2.31	2.15	1.78	**1.69**
Kitti05	7.23	4.33	3.53	**3.20**
Kitti06	11.53	12.37	11.51	**11.12**
Kitti07	10.38	**3.66**	4.32	4.24
Kitti08	25.34	11.57	8.89	**8.71**
Kitti09	11.54	8.57	**7.26**	7.29
Kitti10	9.58	8.34	6.85	**6.64**
V101	0.85	0.51	0.33	**0.29**
V102	1.26	0.73	0.42	**0.39**
V103	1.83	1.55	1.41	**1.38**
V201	0.51	0.59	0.43	**0.38**
V202	0.97	0.86	0.50	**0.45**
V203	X	X	2.15	**1.93**

Bold value indicates the coresponding method achieves best results

were compared with those of the algorithms above. The numbers in bold indicate the best results among the four methods. The proposed PLWM-VO algorithm achieved the best performance on the test sequences, except for Kitti07. PL-SVO demonstrated a significant error accumulation owing to the frame-to-frame pose-estimation method. Compared with the ORB-VO algorithm, the proposed method reduced the ATE for the trajectory by an average of 17.88% in the Kitti dataset, which corresponds to an average reduction of 36.79% in Europe. These results indicate that the proposed algorithm offers significant improvement in terms of pose estimation. Second, the ATEs of PLWM-VO and PLWM-VO-without-SAD were compared. In the Kitti dataset, the ATE for the trajectory generated by PLWM-VO reduced by 3.30% on average compared with that generated by PLWM-VO-without-SAD. In the Euroc dataset, that is reduced by an average of 8.87%. In the Euroc dataset, that is reduced by an average of 8.87%. This shows that accurately detecting endpoints using the SAD algorithm can improve the accuracy of pose estimation. Scenes with unevenly distributed features were ubiquitous in the Euroc and Kitti datasets. The results of the method comparison show that the location accuracy of the proposed algorithm improved in different degrees in the Euroc and Kitti datasets. This indicates that the environmental adaptability of our algorithm improved in environments with uniform and unevenly distributed feature distributions.

To demonstrate the superiority of the PLWM-VO algorithm, the trajectories corresponding to PLWM-VO, ORB-VO, and PL-SVO in the two Kitti02, v101 datasets are compared, as shown in Figures 10 and 11. Red represents the ground truth, and blue indicates the trajectories of the three algorithms. The proposed method performed better than the other two algorithms. To render the experiment more convincing, some areas in the trajectory (within the area of the black dotted rectangle) were aligned, and a comparison was performed with the ground truth. As shown in Figure 9, the ATEs of the partial trajectory generated by the PL-SVO, ORB-VO, and PLWM-VO algorithms in Kitti 02 are 18.706, 9.007, and 4.663 m, respectively. Meanwhile, as shown in Figure 10, the ATEs of the partial trajectory generated by the PL-SVO, ORB-VO, and PLWM-VO algorithms in v101 are 0.178, 0.146, and 0.073 m, respectively. Therefore, the proposed algorithm demonstrates good pose estimation performance.

Additionally, the average operating time of each frame in the Kitti and Euroc datasets was obtained (see Table 3). The result shows that the proposed algorithm is more time consuming than the ORB-VO and PL-SVO because of the addition of SAD line endpoint detection and the weight-adaptive model. Although inefficient, it can satisfy the requirements of real-time operations.

3.2 Online Tests

Online tests were performed on the proposed PLWM-VO method in three environments using a stereo camera (MYNT EYE S1030-IR, whose baseline and resolution are 120 mm and 640 pixels × 480 pixels, respectively). The laboratory represents an environment with numerous points and lines. A building room represents an

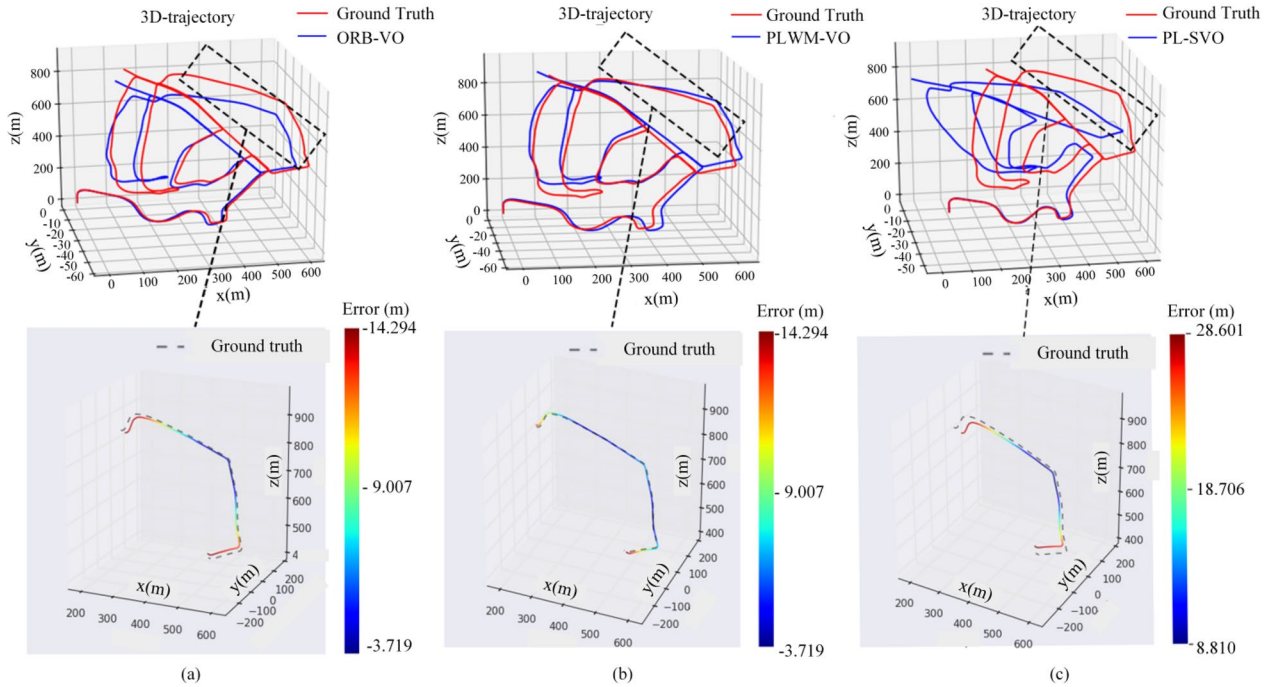

Figure 10 Comparisons of heat maps obtained using ORB-VO, PL-SVO, and the proposed method in Kitti 02: **a** ORB-VO, **b** PLWM-VO, and **c** PL-SVO (blue and red lines correspond to lower and higher error levels, respectively)

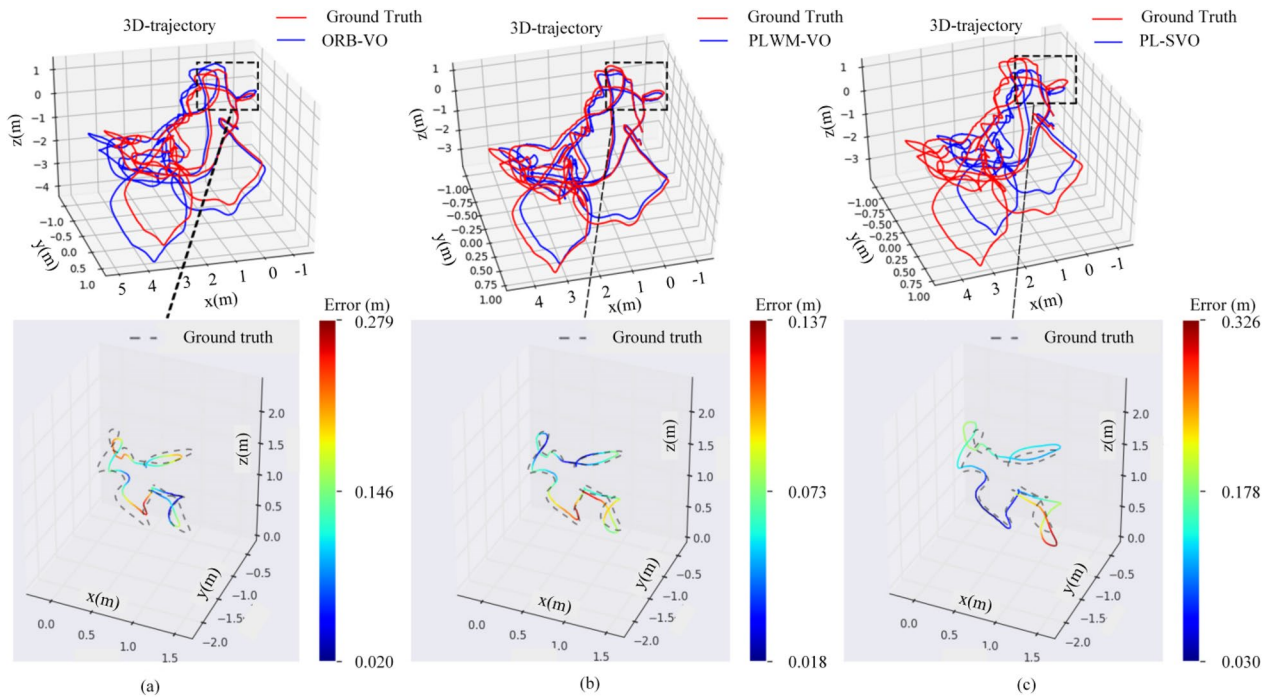

Figure 11 Comparisons of heat maps obtained using ORB-VO, PL-SVO, and the proposed method in sequence V101: **a** ORB-VO, **b** PLWM-VO, and **c** PL-SVO (blue and red lines correspond to lower and higher error levels, respectively)

Table 3 Average time of four methods per frame

Dataset	Processing time (ms)			
	PL-SVO	ORB-VO	PLWM-VO-without SAD	PLWM-VO
Kitti	133.59	116.85	146.89	162.46
Euroc	95.04	86.73	110.21	116.47

artificial environment with low-texture scenes, whereas a campus represents a point-rich environment.

The ground truth is difficult to obtain during the online tests. Therefore, in the tracking process, the stereo camera begins tracking at an initial point, turns in a circle in the target area, and then returns to the origin. The distance between the endpoint and the origin of the estimated trajectory is measured to obtain the cumulative translation error. The specific calculation process is as follows: The movement of the experimental equipment was controlled during the experiment; subsequently, the equipment was shifted to the starting point. The offset distance between the starting and end points in the trajectory was drawn via measurement; subsequently they were divided by the total distance of the movement to obtain the cumulative translation error.

The three environments and the corresponding trajectories generated by the proposed algorithm are shown in Figure 12. Top figures show characteristics of three environments, bottom figures show trajectories corresponding to the three environments, and red arrows indicate direction of movement. Table 4 shows that the cumulative translation errors in all three environments are less than 0.8% of the total distance. The PLWM-VO algorithm not only exhibited good performance in point-rich environments (campus), but also performed well in point-line-rich environments (the laboratory) and artificial environments with low-texture scenes (building room). Thus, it can satisfy the usage requirements in various environments.

4 Discussion

The aim of this study was to improve the environmental adaptability of visual odometry for scenes with unevenly distributed features, which are ubiquitous in visual odometry. Current algorithms can easily yield weak-point

Table 4 Translation error in three different environments

Experiment scene	Distance (m)	Translation error (m)	Relative error (%)
Laboratory	32.68	0.21	0.64
Campus	235.16	1.81	0.77
Building room	104.77	0.69	0.66

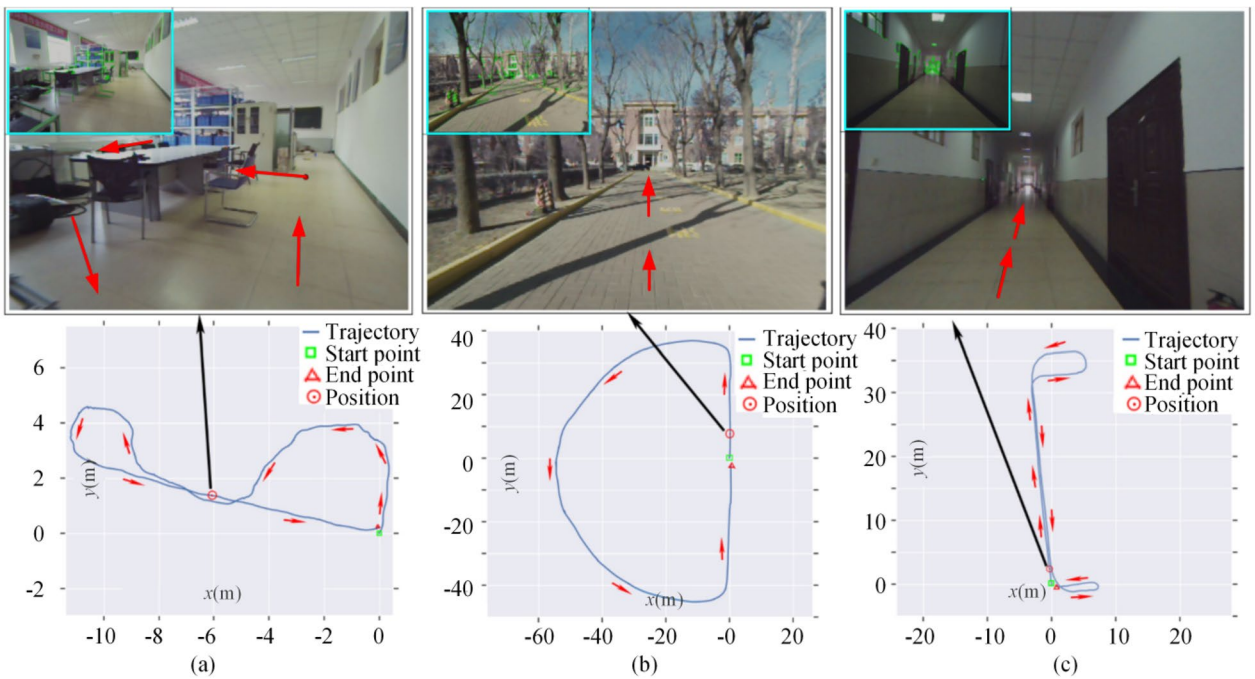

Figure 12 Trajectories estimated in three environments: **a** Laboratory, **b** campus, and **c** building room

features. Moreover, they are unsuitable for environments with low-texture scenes. The human visual attention mechanism offers good adaptability to scenes with uneven feature distributions; it is based on a global and balanced approach for scene understanding. Furthermore, it prioritizes distinctive objects with specific features (color, location, etc.).

Herein, we propose a weight-adaptive model based on region partition and region growth to simulate the human visual attention mechanism in visual odometry. Figure 7 shows that as the region division parameter n changed from 1 to 9, the weight differences between the dense and sparse regions reduced significantly (from 28.1%, 31.6%, and 34.8% to 5.4%, 15.0%, and 11.3%, respectively). This indicates that the proposed PLWM-VO can improve the globality and balance of feature descriptions by optimizing the weight distribution between the feature-sparse and -dense regions. Furthermore, as shown in Table 2, compared with ORB-VO, our algorithm improved the location accuracy for the Kitti and Euroc datasets by 17.88% and 36.79% on average, respectively. This indicates that the environmental adaptability of visual odometry can be improved in scenes with unevenly distributed features by introducing a human visual attention mechanism.

Although this study primarily investigates the effect of the human visual attention mechanism on visual odometry, particularly on its adaptability to a scene with unevenly distributed features, it provides a foundation for improving SLAM and robot navigation. The improvement in pose estimation accuracy provides a basis for improving mapping during SLAM and robot navigation in different environments. Furthermore, the mechanism can be extended to other robotic fields such as autonomous driving, social service robots, security robots, and drones. In future studies, we will attempt to extend the human visual attention mechanism to location and mapping in SLAM and to the navigation technology of robots.

Furthermore, in the physiology field, we plan to provide a demonstration and a quantitative approach for investigating the effectiveness of the human visual attention mechanism in visual information processing. Sacrey et al. [34] investigated the relationship between the human visual attention mechanism and human behavior. The accuracy of grasping items was used as a quantitative evaluation index to verify the relationship between visual guidance and movement coordination. This study demonstrated that the position estimation accuracy of visual odometry can be adopted as a quantitative evaluation index to further investigate the mechanism of human visual attention.

5 Conclusions

To improve the adaptability of visual odometry to environments with unevenly distributed features, we proposed a human visual attention mechanism-inspired point-and-line stereo visual odometry algorithm known as PLWM-VO. The effectiveness of the proposed algorithm was verified using experimental results.

(1) By developing a weight-adaptive model based on region partition and region growth, we implemented a human visual attention mechanism to rationally assign weights between dense and sparse features. Thus, the globality and balance of the description of unevenly distributed features improved significantly.
(2) A modified reprojection error model was established using the SAD algorithm to improve the detection accuracy of line endpoints. The results showed that the introduction of SAD further improved the pose-estimation accuracy of PLWM-VO.
(3) Compared with the state-of-the-art method (ORB-VO), the proposed PLWM-VO improved the position estimation accuracy by 36.79% in environments with unevenly distributed features, based on the overall results of open datasets and online tests.

The proposed method yielded good results in general scenarios and better positioning accuracy. However, owing to its unstable line characteristics, it poses challenges when used in scenes with significant light changes. In the future, features can be extracted based on deep learning or the conventional edge detection, which can overcome the problem of unstable feature detection in scenes with significant light changes.

Acknowledgements
Not applicable.

Author contributions
CW directed the studies and was a major contributor in writing the manuscript; JZ was responsible for the entire research; YZ directed the studies and was a major contributor in writing the manuscript; YJZ performed the experiments and assisted with the analysis of the results; JJ programmed the algorithm. All authors read and approved the final manuscript.

Authors' Information
Chang Wang was born in 1988. He received his B.E., M.E. and Ph.D. degrees in mechanical engineering from *Yanshan University, China,* in 2011, 2014 and 2021, respectively. He is currently a lecturer at *School of Mechanical Engineering, University of Science and Technology Beijing*. His current research interests include intelligent robots, control and mechanical engineering.
Jianhua Zhang was born in 1979. He received his B.E. degree in Mechanical Engineering from *Agricultural University of Hebei, China*, in 2003, and his M.E. and Ph.D. degrees in mechanical engineering from *Shandong University,*

China in 2005 and 2008, respectively. He is currently a professor at *School of Mechanical Engineering*, *University of Science and Technology Beijing*. His research interests include intelligent robots, rehabilitation exoskeletons and collaborative robots.

Yan Zhao was born in 1988. He received his B.E. and M.E. degrees in mechanical engineering from *Hebei University of Technology, China,* in 2012 and 2015, respectively, and his Ph.D. degree in biomedical engineering from *Beijing Institute of Technology, China*, in 2019. He is currently a lecturer at *School of Mechanical Engineering*, *University of Science and Technology Beijing*. His current research interests include surgical robotics, medical image processing, deep learning, and robotic learning.

Youjie Zhou was born in 1994. He received his B.E. and M.E. in mechanical engineering from *Qingdao University of Technology, China*, in 2017, and his M.E. in mechanical engineering from *Hebei University of Technology, China*, in 2020. He is currently pursuing a Ph.D. degree in mechanical engineering at *Shandong University, China*. His current research interests include visual SLAM and positioning.

Jincheng Jiang was born in 1998. He is currently pursuing a master's degree in mechanical engineering at *School of Mechanical Engineering, Hebei University of Technology, China*. He received a bachelor's degree in mechanical design, manufacturing, and automation from *Yangtze Normal University, China*. His current research interests include visual SLAM and machine learning.

Funding

Supported by Tianjin Municipal Natural Science Foundation of China (Grant No. 19JCJQJC61600), Hebei Provincial Natural Science Foundation of China (Grant Nos. F2020202051, F2020202053).

Declarations

Competing Interests

The authors declare no competing financial interests.

Received: 14 August 2021 Revised: 25 February 2023 Accepted: 7 March 2023
Published online: 09 May 2023

References

[1] S Wang, R Clark, H Wen, et al. Deepvo: Towards end-to-end visual odometry with deep recurrent convolutional neural networks. *Proceedings of the IEEE International Conference on Robotics and Automation*, Marina Bay, Singapore, May 29-June 3, 2017: 2043-2050.

[2] V Guizilini, F Ramos. Visual odometry learning for unmanned aerial vehicles. *Proceedings of the IEEE International Conference on Robotics and Automation*, Shanghai, China, May 9-13, 2011: 6213-6220.

[3] A Handa, T Whelan, J McDonald, et al. A benchmark for RGB-D visual odometry, 3D reconstruction and SLAM. *Proceedings of the IEEE International Conference on Robotics and Automation*, Hong Kong, China, May 31-June 5, 2014: 1524-1531.

[4] A Bera, T Randhavane, E Kubin, et al. The socially invisible robot navigation in the social world using robot entitativity. *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, Madrid, Spain, October 1-5, 2018: 4468-4475.

[5] F Han, H Wang, G Huang, et al. Sequence-based sparse optimization methods for long-term loop closure detection in visual SLAM. *Autonomous Robots*, 2018, 42(7): 1323-1335.

[6] A Pumarola, A Vakhitov, A Agudo, et al. PL-SLAM: Real-time monocular visual SLAM with points and lines. *Proceedings of the IEEE International Conference on Robotics and Automation*, Marina Bay, Singapore, May 29-June 3, 2017: 4503-4508.

[7] E Perdices, L M López, J M Canas. LineSLAM: Visual real time localization using lines and UKF. *Proceedings of the ROBOT2013: First Iberian Robotics Conference*, Madrid, Spain, November 28-29, 2013: 663-678.

[8] M Hsiao, E Westman, G Zhang, et al. Keyframe-based dense planar SLAM. *Proceedings of the IEEE International Conference on Robotics and Automation*, Marina Bay, Singapore, May 29-June 3, 2017: 5110-5117.

[9] Y Ling, M Kuse, S Shen. Edge alignment-based visual–inertial fusion for tracking of aggressive motions. *Autonomous Robots*, 2018, 42(3): 513-528.

[10] P Delmas, T Gee. Stereo camera visual odometry for moving urban environments. *Integrated Computer-Aided Engineering*, 2019, 26(3): 243-256.

[11] A J Davison, I D Reid, N D Molton, et al. MonoSLAM: Real-time single camera SLAM. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2007, 29(6): 1052-1067.

[12] G Klein, D Murray. Parallel tracking and mapping for small AR workspaces. *Proceedings of the IEEE and ACM International Symposium on Mixed and Augmented Reality*, Nara, Japan, November 13-16, 2007: 225-234.

[13] P Henry, M Krainin, E Herbst, et al. RGB-D mapping: Using depth cameras for dense 3D modeling of indoor environments. *Proceedings of the Experimental Robotics,* Springer, Berlin, Heidelberg, 2014: 477-491.

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

[15] E Rublee, V Rabaud, K Konolige, et al. ORB: An efficient alternative to SIFT or SURF. *Proceedings of the International Conference on Computer Vision*, Barcelona, Spain, November 6-13, 2011: 2564-2571.

[16] J Witt, U Weltin. Robust stereo visual odometry using iterative closest multiple lines. *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, Tokyo, Japan, November 3-8, 2013: 4164-4171.

[17] Y Lu, D Song. Robust RGB-D odometry using point and line features. *Proceedings of the IEEE International Conference on Computer Vision*, Santiago, Chile, December 7-13, 2015: 3934-3942.

[18] S J Li, B Ren, Y Liu, et al. Direct line guidance odometry. *Proceedings of the IEEE International Conference on Robotics and Automation*, Brisbane, Australia, May 21-25, 2018: 5137-5143.

[19] R Gomez-Ojeda, J Gonzalez-Jimenez. Robust stereo visual odometry through a probabilistic combination of points and line segments. *Proceedings of the IEEE International Conference on Robotics and Automation*, Stockholm, Sweden, May 16-21, 2016: 2521-2526.

[20] R Gomez-Ojeda R, F A Moreno, D Zuniga-Noël, et al. PL-SLAM: A stereo SLAM system through the combination of points and line segments. *IEEE Transactions on Robotics*, 2019, 35(3): 734-746.

[21] R G Von Gioi, J Jakubowicz, J M Morel, et al. LSD: A fast line segment detector with a false detection control. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2008, 32(4): 722-732.

[22] X Zuo, X Xie, Y Liu, et al. Robust visual SLAM with point and line features. *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, Vancouver, BC, Canada, September 24-28, 2017: 1775-1782.

[23] R Mur-Artal, J D Tardós. ORB-SLAM2: An open-source slam system for monocular, stereo, and RGB-D cameras. *IEEE Transactions on Robotics*, 2017, 33(5): 1255-1262.

[24] M Jian, K M Lam, J Dong, et al. Visual-patch-attention-aware saliency detection. *IEEE Transactions on Cybernetics*, 2014, 45(8): 1575-1586.

[25] M W Guy, G D Reynolds, D Zhang. Visual attention to global and local stimulus properties in 6-month-old infants: Individual differences and event-related potentials. *Child Development*, 2013, 84(4): 1392-1406.

[26] A Romberg, Y Zhang, B Newman, et al. Global and local statistical regularities control visual attention to object sequences. *Proceedings of the Joint IEEE International Conference on Development and Learning and Epigenetic Robotics*, Paris, France, September 19-December 22, 2016: 262-267.

[27] J Vanne, E Aho, T D Hamalainen, et al. A high-performance sum of absolute difference implementation for motion estimation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2006, 16(7): 876-883.

[28] L Zhang, R Koch. An efficient and robust line segment matching approach based on LBD descriptor and pairwise geometric consistency. *Journal of Visual Communication and Image Representation*, 2013, 24(7): 794-805.

[29] A Sengupta, S Elanattil. New feature detection mechanism for extended Kalman filter based monocular SLAM with 1-Point RANSAC. *Proceedings of the International Conference on Mining Intelligence and Knowledge Exploration*, Hyderabad, India, December 9-11, 2015: 29-36.

[30] Y Liu, Y Gu, J Li, et al. Robust stereo visual odometry using improved RANSAC-based methods for mobile robot localization. *Sensors*, 2017, 17(10): 2339.

[31] A Geiger, P Lenz, R Urtasun. Are we ready for autonomous driving? The kitti vision benchmark suite. *Proceedings of the IEEE Conference on*

Computer Vision and Pattern Recognition, Providence, Rhode Island, 2012: 3354–3361.
[32] M Burri, J Nikolic, P Gohl, et al. The EuRoC micro aerial vehicle datasets. *The International Journal of Robotics Research*, 2016, 35(10): 1157-1163.
[33] J Sturm, N Engelhard, F Endres, et al. A benchmark for the evaluation of RGB-D SLAM systems. *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, Vilamoura-Algarve, Portugal, October 7-12, 2012: 573-580.
[34] L A R Sacrey, J M Karl, I Q Whishaw. Development of visual and somatosensory attention of the reach-to-eat movement in human infants aged 6 to 12 months. *Experimental Brain Research*, 2012, 223(1): 121-136.