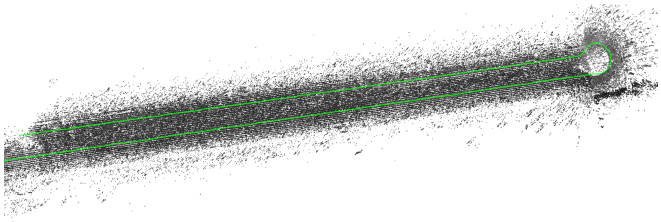
2021 IEEE International Conference on Robotics and Automation (ICRA 2021)
May 31 - June 4, 2021, Xi'an, China

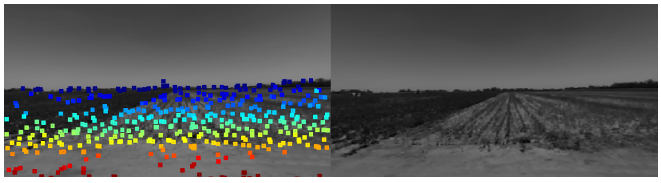
Accurate and Robust Stereo Direct Visual Odometry for Agricultural Environment

Tao Yu[1,2], Junwei Zhou[1], Liangliang Wang[1] and Shengwu Xiong[1,2]

***Abstract*— Vision-based localization and mapping in the agricultural environment is challenging due to the unstructured scene with unstable features, illumination variations, bumpy roads, and dynamic environmental objects. To address these challenges, we propose an accurate and robust stereo direct visual odometry system with modifications on Stereo-DSO. We firstly select some well-matched static stereo points in the latest keyframe to improve the accuracy of inverse depth calculation for tracking. The inverse depth can further distinguish close objects from background, which will avoid large and far-away scene objects in keyframe determination. To boost efficiency and accuracy at the tracking stage, we propose a point selection method to sample map points and remove outliers. Furthermore, altitude smoothness verification with a local flat ground assumption and recovery method for tracking failure on bumpy roads are proposed to improve the system's robustness. Finally, a far-away keyframe is reserved in the sliding window to alleviate the orientation drift since the agricultural robots usually move straightly following the crop row. Our system achieved new state-of-the-art results on Flourish dataset and the recently released Rosario dataset.**



(a) Trajectory and sparse map.



(b) Keyframe and current frame.

Fig. 1. Visual output of our system on Rosario dataset [7].

I. INTRODUCTION

Autonomous agricultural robots have drawn many attention [1] over the past years, exhibiting many potential applications in precision agriculture, such as weeding, crop monitoring, and harvesting. In this context, simultaneous localization and mapping (SLAM), and visual odometry (VO) systems are the foundation of agricultural robots for sensing, perception, and interpretation of the crops. However, localization and mapping in the agricultural scene are challenging due to the unstructured environment with less stable features and more repetitive textures, illumination change, bumpy road, and weakly dynamic environmental objects caused by wind. To this end, a large number of systems are proposed with a high-end Real-Time Kinematic Global Navigation Satellite System (RTK-GNSS) to obtain an accurate position for agricultural robots [2]. However, these sensors are neither too expensive nor requiring extra ground stations. Moreover, the GNSS can only provide position information of robots without a map of the environment. In contrast, vision sensors can collect rich texture information to provide location and map by SLAM algorithms. Besides, they are typically cheap and straightforward for widespread deployment. Vision-based SLAM (vSLAM) has gained more attention in the field of autonomous robotic.

In recent years, real-time VO and vSLAM have made significant progress in a range of wild applications, such as autonomous driving, augmented reality, and robotic navigation. Some vSLAM systems have been proposed based on indirect(including feature-based and optical flow-based method), such as ORB-SLAM2 [3], VINS-Fusion [4]. Some vSLAM systems are based on direct methods, such as Stereo-DSO [5]. However, most researches focus on indoor or urban environments with obvious structural features. Unfortunately, these algorithms can not perform well in a challenging agricultural environment. It is difficult for feature-based methods to track with fewer distinguishable key points in the agricultural scene. Moreover, the traditional place recognition algorithms based on Bag of Word (BoW) [6] usually failed in this environment, resulting in no loop detected for global consistency. On the other hand, optical flow-based methods are more efficient since no complex descriptors of feature points are needed, and features are tracked by the raw intensity of the image. Unfortunately, such methods are not accurate enough for a long-time motion. Direct methods could provide accurate and robust motion estimation owing to the full usage of image information. However, non-convex energy function based optimization highly depends on the initial values of pose and depth. Tracking failures may happen due to the motion on bumpy road against the constant motion assumption.

In this paper, to address the agricultural environment's challenges, we propose an accurate and robust stereo direct VO system based on Stereo-DSO [5]. Fig. 1 shows original stereo images and maps with trajectory output by our system. The contribution of this work is as follows: 1) To the best of our knowledge, this is the first work evaluating stereo direct

[1]School of Computer, Wuhan University of Technology, Wuhan, 430070
[2]Sanya Science and Education Innovation Park of Wuhan University of Technology, Sanya 572000

visual odometry in agricultural environments. 2) An accurate and robust stereo direct VO system with a set of strategies is proposed for the challenging agricultural scene. 3) State-of-the-art results are obtained on Flourish [8] dataset and recently released Rosario [7] dataset.

II. Related Work

1) SLAM/VO Systems: SLAM and VO systems have been a popular topic over the past decades. Extensive visual-based algorithms on location and mapping have been proposed, which can be generally divided into two categories, indirect and direct methods. For indirect methods, a certain feature(e.g., ORB [9], SURF [10]) points are extracted from images and tracked by feature matching or optical flow. In recent years, many systems [11] [12] [3] [4] have been proposed to improve the accuracy and robustness of location and mapping. Unlike indirect methods, direct methods aim at estimating camera motion and 3D environment structure by using the original intensities of the image. Some notable systems such as DTAM [13], LSD-SLAM [14], SVO [15] have been developed to obtain accurate motion estimation and ease the computation cost. More recently, DSO [16] has been a state-of-the-art VO system in terms of both accuracy and robustness on a large monocular camera tracking dataset [17]. Stereo-DSO [5] was an extension of DSO [16] to achieve highly accurate real-time visual odometry estimation of large-scale environments from a stereo camera.

2) Long Range Stereo: In the agricultural scene, stereo vision faces difficulty since the distant view. To obtain better stereo range resolution, Gabe et al. [18] proposed a novel iterated sigma point Kalman filter for the long-range stereo problem. It achieved superior performance in terms of efficiency and accuracy. Peter et al. [19] performed a comprehensive statistical evaluation of some state-of-the-art stereo matching approaches and presented guidelines on algorithmic choices derived from theory. Pasquale et al. [20] performed a comparison between monocular and stereo vision systems for long-range pose estimation.

3) Localization and Mapping in Agricultural Environments: To address the challenges, it is reasonable to utilize the characteristics of plants [21] [22] [23] [24]. A set of researches [25] [26] [27] [28] [29] focus on crop detection and following for robot navigation in the agricultural scene. However, these methods have limited application scope due to their dependence on plants in the field. More recently, to improve SLAM's accuracy and robustness, multi-sensor fusion and collaborative mapping systems have been proposed. Imperoli [8] designed a multi-sensors(including GNSS receiver, camera, IMU) positioning system with Digital Evaluation Map (DEM) for agricultural robots. Dong et al. [30], and Chebrolu et al. [31] fused information from several on-board sensors and aerial maps in a localization system to address the challenge of changing the appearance of the crop fields. Aerial-ground collaborative 3D mapping systems [32] [33] [34] have been proposed for precision agriculture. Most of these systems focus on data fusion approaches rather than a single sensor's performance and require either additional devices or reference maps. Moreover, as we know, accurate and robust estimation by a single sensor could improve the whole localization and mapping system's performance.

4) Evaluation of Visual SLAM in Agricultural Environments: To our knowledge, only two evaluations of visual SLAM in agricultural environments are conducted. The first [35] evaluated visual SLAM strategies for stereo cameras applied to the trajectory estimation on a self-propelled platform in a typical fruit environment. S-PTAM, ORB-SLAM2, LibViso are used for assessment. In the second work [36], S-PTAM [12], ORB-SLAM2 [3], S-MSCKF [37] are assessed on Rosario dataset [7] and achieve poor performance on accuracy and robustness compared with performances reported on urban or indoor environments. However, all the introduced evaluations are indirect without comparisons with any state-of-the-art direct methods which achieve good performance on the public VO/SLAM datasets [38] [39] [40].

III. Accurate and Robust Stereo Direct Visual Odometry

In our stereo direct visual odometry system, we use the same strategy as Stereo-DSO [5] to initialize the whole system. Static stereo matching algorithm is utilized to generate a semi-dense map. We randomly select N_f points to be tracked in the next frame. Then we use direct image alignment to track new frames with the constant motion assumption. The failure recovery process will start if tracking failure happens. Finally, the joint window optimization to improve system's accuracy is performed in another thread.

A. Direct Image Alignment

Consider the $i-th$ reference frame F_i, associated with a gray-scale image I_i and a point set P_i with the inverse depth of each point. P_i in F_i is observed by another frame F_j. The energy of direct image alignment in these two frames can be formulated as

$$\begin{aligned} E_{ij} &= \sum_{\mathbf{p}\in\mathbf{P}_i} \omega_{\mathbf{p}} \left\| r_{ij}(\mathbf{p}) \right\|_{\gamma} \\ r_{ij}(\mathbf{p}) &= I_j(\mathbf{p}^{'}) - b_j - \frac{e^{a_j}}{e^{a_i}}(I_i(\mathbf{p}) - b_i) \end{aligned} \tag{1}$$

where $a_i,\ b_i,\ a_j,\ b_j$ denote the affine brightness parameters with transfer function given by $e^{-a_i}(I_i - b_i)$, while $\|\cdot\|_{\gamma}$ is the Huber norm and the $\omega_{\mathbf{p}}$ is a weighting to down-weight the point with high image gradient.

$$\omega_{\mathbf{p}} = \frac{\varepsilon^2}{\varepsilon^2 + \|\nabla I_i(\mathbf{p})\|_2^2} \tag{2}$$

where ε is a constant for numerical stability. $\mathbf{p}^{'}$ is projection of $\mathbf{p}$ from I_i to I_j calculated by

$$\mathbf{p}^{'} = \pi(T_{ji}\pi^{-1}(\mathbf{p},\ d_{\mathbf{p}})) \tag{3}$$

$d_{\mathbf{p}}$ represents the inverse depth of $\mathbf{p}$ and $\pi(\cdot)$ and $\pi^{-1}(\cdot)$ are camera projection and back-projection functions, while T_{ji} denotes the 3D rigid body transformation from F_i to F_j

$$T_{ji} = \begin{bmatrix} R_{ji} & t \\ 0 & 1 \end{bmatrix} = T_j^{-1}T_i,\ R_{ji} \in SO(3),\ t \in \mathbb{R}^3 \tag{4}$$

B. Tracking in Agricultural Scene

In this step, the system obtains the pose of the new frame and determines if a keyframe should be created. When a new frame is fed into the system, we calculate the inverse depth of points in the latest keyframe first. Then we sample map points and remove outliers by the proposed method. Finally, we select close tracked points for keyframe determination.

Inverse Depth Calculation for the Latest Keyframe: All the activated points in the sliding window are projected onto the latest keyframe. Some activated points $\mathbf{P}_p$ are mapped into the same position. A weighted average inverse depth $\overline{id}$ of the points is calculated by

$$\overline{id} = \sum_{i \in \mathbf{P}_p} (\omega_i * id_i) / \sum_{i \in \mathbf{P}_p} \omega_i \tag{5}$$

Where ω_i is a weight depending on the uncertainty of inverse depth id_i.

We select some candidate points in the latest keyframe through the quality of static stereo matching to increase the number of tracked points and improve the accuracy of inverse depth calculation. The quality q is defined as

$$E_s(\mathbf{p}_r) = \sum_{\mathbf{p}_1 \in pat(\mathbf{p}_l),\ \mathbf{p}_2 \in pat(\mathbf{p}_r)} \left\| I_r(\mathbf{p}_2) - b_r - \frac{e^{a_r}}{e^{a_l}} (I_l(\mathbf{p}_1) - b_l) \right\|_\gamma$$
$$q = second \min_{\mathbf{p}_r \in \mathbf{P}_e} E_s(\mathbf{p}_r) / \min_{\mathbf{p}_r \in \mathbf{P}_e} E_s(\mathbf{p}_r) \tag{6}$$

Where $\mathbf{P}_e$ is a set of point on the epipolar line of right image I_r associated with point $\mathbf{p}_l$ on the left image I_l. a_l, a_r, b_l, b_r are the affine brightness parameters of left and right images. $pat(\mathbf{p})$ is an 8-point patch of $\mathbf{p}$, while E_s represents the energy of image patch's similarity. Besides, we adopt the depth limitation defined in (7) to avoid bringing far-away static stereo points with large errors. q is assigned to the weight of the inverse depth to keep consistency with (5).

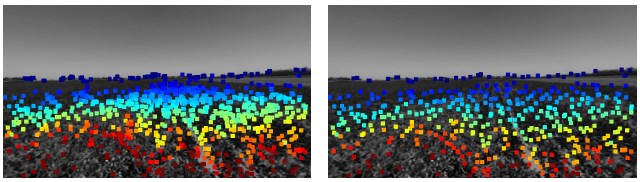
$$d < \theta * b \tag{7}$$

Where d is the depth of static stereo points and b is the baseline of camera. θ is a threshold to filter far-away points.

Point Selection: The non-uniform distributed points with some outliers are obtained from inverse depth calculation, which can introduce errors to the optimization. To address this problem, we proposed a selection method to sample points and remove outliers as follows: 1) split the image into $grid \times grid$ blocks. 2) calculate the mean d_{mean} and standard deviation d_{std} of depth for each block. 3) sort points according to their weights. 4) reserve only one point with max weight and its depth d subject to $|d - d_{min}| < 3d_{std}$, and remove other points in this block. An example of this method is shown in Fig. 2. Fewer points can increase the tracking stage's efficiency. Moreover, the tracking accuracy can be improved since some outliers are removed.

Pose Optimization: We optimize the pose of the new frame by minimizing the energy E in (8) while keeping depth fixed. Gauss-Newton method on image pyramid in a coarse-to-fine order is performed.

$$E = \sum_{\mathbf{p} \in \mathbf{P}_i} (\omega_\mathbf{p}^L \left\| r_{ij}^L(\mathbf{p}) \right\|_\gamma + \omega_\mathbf{p}^R \left\| r_{ij}^R(\mathbf{p}) \right\|_\gamma) \tag{8}$$



(a) Original points. (b) Selected points.

Fig. 2. An example of point selection.

Where r_{ij}^L is the left-to-left photometric error, while r_{ij}^R denotes the photometric error from the left image of the keyframe to the right image of the current frame. We skip the point whose projection from the reference keyframe to the new frame is out of the image boundary.

Keyframe Criteria: For agricultural scenes, since the wide view of vision, large and far-away objects such as trees will always be visible in each frame. Therefore, few keyframes can be created with keyframe criteria in [16], and the marginalization always happens in new keyframes, resulting in frames with large translations remaining in the sliding window. Moreover, most activated points are from far large objects. It introduces significant scale errors in tracking result, which corrupts the system. We change the keyframe criteria from all points to specific points whose depths are less than θ times motion distance to avoid this situation.

C. Tracking Failure Recovery

In the agricultural environment, most of the visual features are extracted from ground planes. Hence, the measurement of pixels with small errors can cause non-negligible drift along the vertical direction. On the other side, typical agricultural field presents locally flat ground levels, which means the robots' motions along the vertical direction should be small. Motivated by this observation, we verify the smoothness along the vertical direction to avoid potential incorrect tracking. The local trajectories on the yz-plane should approximate a line. Therefore, the criteria can be formulated as

$$d_z = \frac{|t_z - k t_y|}{\sqrt{1+k^2}},\ k = \sum_{t_i \in T} t_{iz} / \sum_{t_i \in T} t_{iy} \tag{9}$$

Where $(t_x,\ t_y,\ t_z)^T$ indicates the translation of consecutive frames and T is a set of recent translations. The direction of xyz-axis is rightward, forward, and upward, respectively. If the translation along z-axis d_z is larger than the threshold, the current frame will be tracked with the motion assumption in (10) by another thread. The failure recovery process will start if an obvious difference between the two poses presents.

Constant motion assumption is usually simple and efficient at the tracking stage. However, in this off-road scene, road's roughness results in the vertical vibration of the robot. Hence tracking with the constant motion model may sometimes fail since the large initial error, which happened in our experiments. To recover from this situation, we make a motion assumption with opposite rotation around x-axis and zero translation along the vertical direction when the constant

motion assumption fails. The new motion assumption $(R^{'},\ t^{'})$ can be obtained by

$$\begin{aligned}(\alpha,\ \beta,\ \gamma) &= f(R)\\ R^{'} &= f^{-1}(-\alpha,\ \beta,\ \gamma)\\ t^{'} &= (t_x,\ t_y,\ k*t_y)^T\end{aligned} \tag{10}$$

Where f is a function that converts a rotation matrix to Euler angles while f^{-1} reverses the operation. $(\alpha,\ \beta,\ \gamma)$ are the rotation angles around the fixed *xyz*-axis, respectively. KLT spare optical flow algorithm [41] is applied to calculate the initial pose of the failure frame with our new motion assumption. Available initial pose of the failure frame can be obtained by the method described in Subsection III-B.

D. Joint Window Optimization

In this step, both the inverse depth of active points and pose of keyframes are jointly optimized to improve the system's accuracy. Assume an active point set $\mathbf{P}$, and $obs(\mathbf{p})$ is a collection of keyframes in the sliding window that can observe the point $\mathbf{p}$. The total energy can be formulated as

$$E_{total} = \sum_{i\in F}\sum_{\mathbf{p}\in\mathbf{P}_i}\Big(\sum_{j\in obs(\mathbf{p})} E_{ij} + \lambda E_{is}\Big) \tag{11}$$

Where E_{ij} denotes the energy of multi-view stereo with left images. E_{is} represents the residuals of static one-view stereo, which is similar to E_{ij} except the fixed transformation obtained from stereo calibration. λ is a balance factor between temporal and static stereo. We use the Gauss-Newton algorithm to optimize pose, affine brightness, and inverse depth parameters.

In the agricultural field, a farming robot usually moves straightly by following the crop rows. The rotation error on the direction of motion could be accumulated in long straight movement and results in an orientation drift. We reserve the oldest keyframe that can observe n_t points in all active keyframes. With this simple strategy, an orientation constraint is constructed by the oldest keyframe and points to alleviate the drift of orientation.

IV. Evaluation

A. Dataset and Metrics

Few datasets are released for localization and mapping in the agricultural scene. The recently public Rosario dataset [7] is multi-sensors data for autonomous robotics in agricultural scenes. It consists of 6 sequences recorded in soybean fields showing real and challenging cases: highly repetitive scenes, reflection, and burned images caused by direct sunlight and rough terrain. Localization and mapping in this environment have become a challenging problem. As described in [7], the difficulty for sequences from 01 to 06 is increasing. We also evaluate our system on another agricultural dataset, named Flourish [8]. The robot moved slowly and smoothly, and the camera is tilted down to capture more ground information. Long backward motion is also presented in this dataset.

To evaluate our system, we report the quantitative results building upon the location errors, including the mean and Root Mean Square Error (RMSE) of the absolute trajectory error (ATE) defined in [42]. All the experiments are conducted on the same desktop computer with Ubuntu 18.04 LTS, Intel Core™ i7-10700 Processor, and 16GB of RAM.

TABLE I

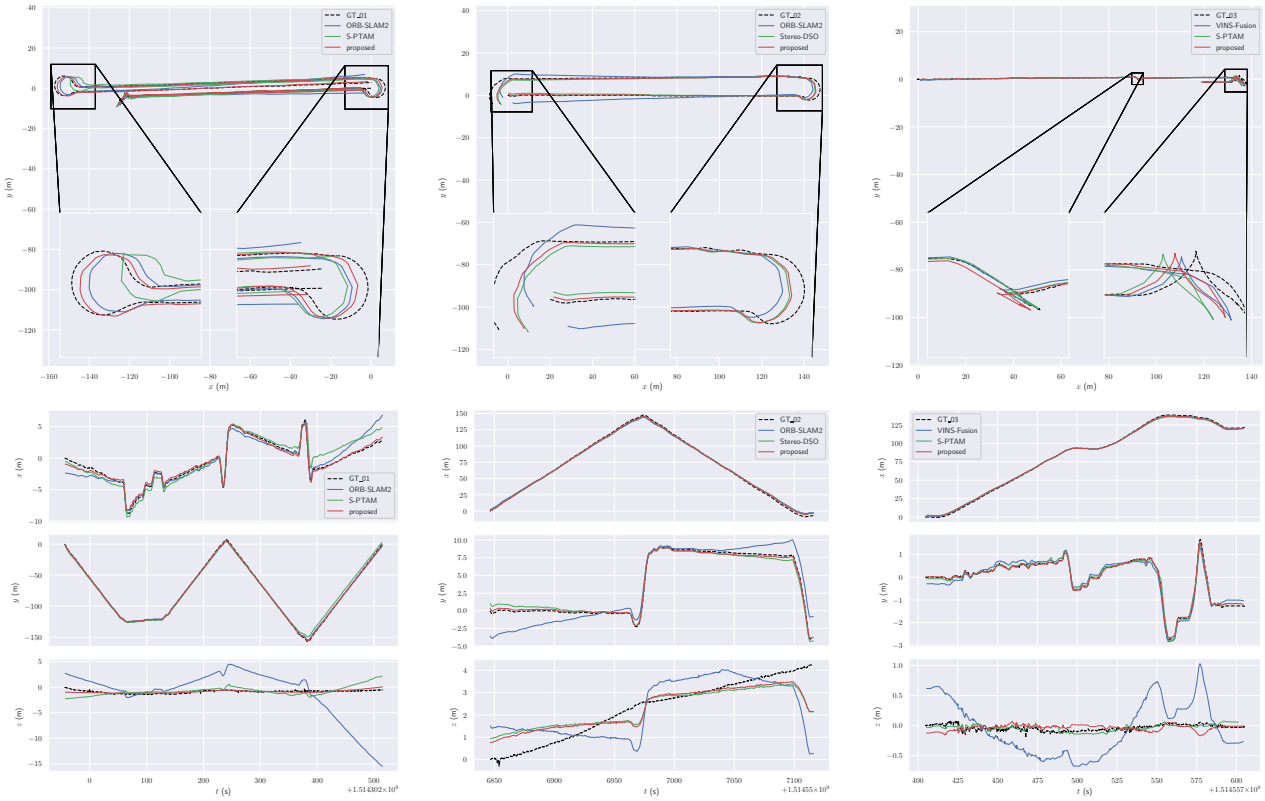
Performance comparisons on Rosario dataset regarding accuracy, robustness and run time.

	Seq	01	02	03	04	05	06
Frames		8624	4160	3092	2542	4839	9091
SVO2	ATE	6.66	8.12	11.42	6.80	×	×
	RMSE	7.36	9.40	12.97	7.95	×	×
S-PTAM	ATE	2.42	2.73	1.43	1.72	×	×
	RMSE	3.14	3.15	1.68	1.90	×	×
ORB-SLAM	ATE	1.89(1)	2.49	3.65	2.53	3.68(4)	5.82(2)
	RMSE	2.20	2.78	4.14	2.92	4.16	6.64
	Time	29/64	26/68	23/47	27/54	26/57	27/55
VINS-Fusion	ATE	5.54	3.64	**0.90**	1.25	2.78	5.18
	RMSE	6.43	4.45	**0.97**	1.36	3.57	6.41
	Time	9/48	9/43	9/43	9/45	9/45	9/51
Stereo-DSO	ATE	6.62	1.73	2.08	1.24	2.57	×
	RMSE	9.20	1.95	2.22	1.43	3.26	×
	Time	13/64	13/46	13/61	12/51	13/57	-
Proposed	ATE	**0.76**	**1.48**	1.00	**1.13**	**0.95**	**2.09**
	RMSE	**0.83**	**1.70**	1.14	**1.30**	**1.10**	**2.34**
	Time	10/40	13/38	11/39	10/39	10/39	10/43
	NSTF	8576	4151	3050	2521	4795	9024

B. System Evaluations

We evaluate our system on Rosario dataset [7] and compare it with state-of-the-art VO/SLAM systems, including SVO2 [15], S-PTAM [12], ORB-SLAM2 [3], VINS-Fusion [4], Stereo-DSO [5]. The overall tracking accuracy in 10 runs is reported in Tab. I. Best results are shown as bold numbers. The number of failure times is presented following the ATE of each sequence if system failures happened. × means the system is always lost in the scene. Besides, the time cost of the last four systems is also shown in Tab. I. Most of these systems are implemented by multi-thread/multi-process. Following ORB-SLAM2 [3], we show mean and max tracking time (*ms*) per frame, which is the most important indicator for real-time performance. We also reported the number of successfully tracked frames (NSTF) with constant motion assumption for our system.

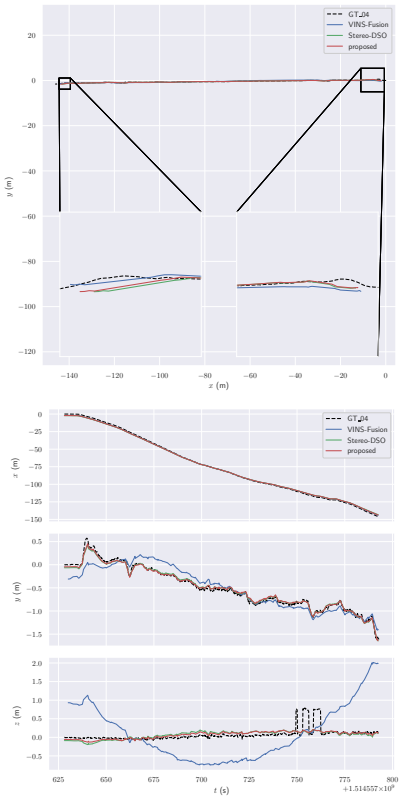
As seen from the Tab. I, our system achieved the best results in all sequences in terms of mean and RMSE ATE except Seq. 03. We obtained a promising 0.83m RMSE in Seq. 01, which is half of the second-best ORB-SLAM2 [3]. Long travel distance and two 180-degree turns are the main difficulties of this sequence. The RMSE of Seq. 02 and 04 are decreased by 0.21m and 0.1m, respectively, compared with the second-best Stereo-DSO. VINS-Fusion [4] achieved promising results in Seq. 03 and 04, while their trajectories are approximately straight. It indicates that the optical flow-based method is applicable in pure straight-line motion. SVO2 [15] achieved poor performance in all sequences. Photometric error without the illumination model is not robust for the agricultural scene. ORB-SLAM2 failed 1, 4, 2 times in Seq. 01, 05, and 06, respectively, due to the agricultural scene's unstable feature points. We obtained remarkable results with 1.14m and 2.43m RMSE in the most



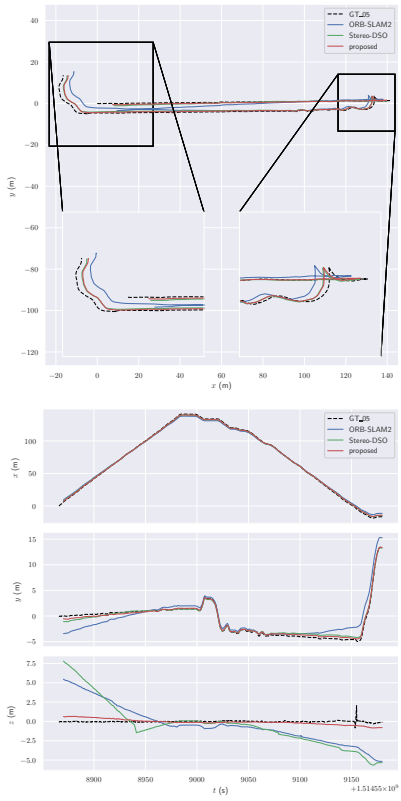
(a) Seq. 01.

(b) Seq. 02.

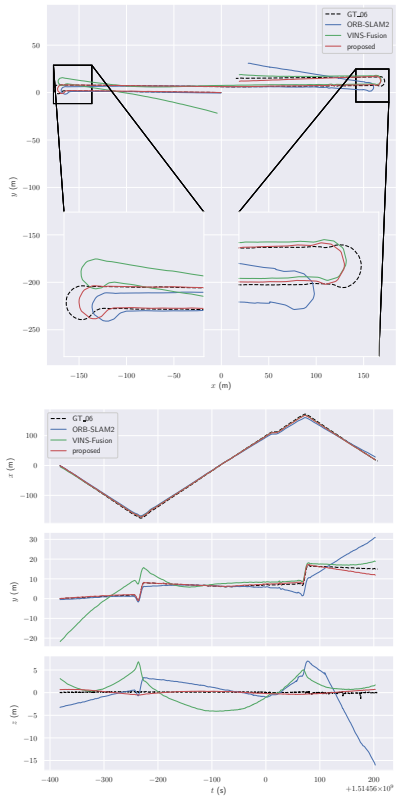
(c) Seq. 03.



(d) Seq. 04.



(e) Seq. 05.



(f) Seq. 06.

Fig. 3. Comparison of the whole trajectories on *xy*-plane and *xyz*-axis obtained from top-3 accurate systems with ground-truth in the six sequences of Rosario dataset.

TABLE II

SYSTEM PERFORMANCE COMPARISON ON FLOURISH DATASET.

	VINS-Fusion	ORB-SLAM2	Stereo-DSO	Proposed
ATE	7.90	1.84	1.67	**1.62**
RMSE	10.31	2.05	1.84	**1.79**
Time(ms)	15/52	30/65	16/58	13/48

difficult Seq. 05 and 06 without system failure. We found that tracking failures happened in each sequence, and our system can recover from the failures in real-time. The max tracking time shows that our system is stable at the tracking stage. The full trajectory results and comparison on different axes of all sequences in the Rosario dataset are shown in Fig. 3. We only keep top-3 accurate results for more clear figures. We can observe that our system's complete trajectory output is more close to the ground-truth than others. Moreover, the most striking observation is the small error on the z-axis output by our system, which improved significantly. In summary, these results show that our system is more accurate and robust in the agricultural environment.

Evaluation on Flourish [8] dataset (15750 frames) is shown in Tab. II. VINS-Fusion [4] performed not well since the backward and long straight motion. The rotation error of the system is accumulated over time. ORB-SLAM2, Stereo-DSO and our system obtained 2.05m, 1.84m, 1.79m RMSE, respectively. We observed that scale error is introduced by the robot configuration and no tracking failure happened since the smooth and slow motion. These experiments indicate that our method is effective in different agricultural scenes.

Failure Analysis: In Seq. 01, we observed that the Stereo-DSO [5] performed a big orientation drift when the robot turned for the last time, resulting in a large total error in trajectory. We found that the bumpy road can decrease the tracking accuracy and even cause tracking failures. If tracking with constant motion assumption fails, relocation by matching the current frame against the keyframe will become more difficult due to the homogeneous scene. Besides, these cases also happened in our system, causing tracking failures. However, the system can perform well with our failure recovery strategy. It indicates that our system is more robust in such challenging conditions common in the agricultural scenes. Other systems can be improved by adopting some similar strategies according to this observation of failures.

C. Ablation Study

To prove the effectiveness of the proposed method, we conducted the following ablation study to understand each step's impact. For the ablation experiments, we adopt the Stereo-DSO as the baseline, and the system is modified with specified setups: a) Keyframe criteria (KC) - using selected points for keyframe determination, b) tracking with some static stereo matching points (TS), c) recovery from failure tracking (RF), d) reserving one far keyframe (RK), and e) applying the point selection method (PR). These configurations are gradually added to the baseline system. The experiment results are shown in Tab. III.

We can observe that the new keyframe criteria achieved slight improvement in easy sequences while significant in difficult Seq. 05, but it still failed in Seq. 06. The system becomes more robust and accurate when adding some static stereo matching points for tracking, even in Seq. 06. With altitude verification and recovery strategy, the system obtained more accurate results, especially in Seq. 01 and 06 that are a long trip with 180-degree turns. The system achieved further improvement in difficult Seq. 05 and 06 with reserving one far keyframe. The slight improvement (0.03-0.12m) of performance was acquired by applying the point selection method. In summary, each step we proposed plays an important role in making the VO system more accurate and robust in the agricultural scene.

TABLE III

MEAN AND RMSE OF THE ABSOLUTE TRAJECTORY ERROR(ATE) FOR THE PROPOSED STRATEGIES AND METHODS ON THE ROSARIO DATASET.

	Seq	01	02	03	04	05	06
Baseline	ATE	6.62	1.73	2.08	1.24	2.57	×
	RMSE	9.20	1.95	2.22	1.43	3.26	×
KC	ATE	5.97	1.54	1.54	1.18	1.35	×
	RMSE	8.51	1.76	1.66	1.36	1.61	×
KC+TS	ATE	2.70	**1.47**	1.17	1.18	1.40	2.59
	RMSE	3.54	**1.68**	1.28	1.36	1.67	3.14
KC+TS+RF	ATE	0.85	1.49	1.07	1.15	1.35	2.29
	RMSE	0.93	1.72	1.21	1.32	1.56	2.54
KC+TS+RF+RK	ATE	0.82	1.51	1.09	1.15	0.98	2.17
	RMSE	0.91	1.73	1.24	1.32	1.14	2.43
All	ATE	**0.76**	1.48	**1.00**	**1.13**	**0.95**	**2.09**
	RMSE	**0.83**	1.70	**1.14**	**1.30**	**1.10**	**2.34**

V. CONCLUSION

In this work, we have proposed an accurate and robust stereo direct visual odometry for the agricultural environment. It achieved new state-of-the-art results in the recent public Rosario dataset [7] and Flourish dataset [8]. More points can be utilized at the tracking stage by adding some static stereo matching points. The point selection method, sampling points and removing outliers, can improve the system's efficiency and accuracy. The failure recovery method makes our VO system more robust for the rough road in the agricultural scene. Using close points to determine the keyframe can ignore large and far-away objects in the agricultural scene. More accurate results are obtained in challenging scenes by reserving one far-away keyframe to constraint the orientation. The experiment results suggest that our system is efficient in challenging agricultural scenes. As to future work, we plan to improve our system's accuracy and robustness in the agricultural environment by fusing visual and inertial measurements.

ACKNOWLEDGMENT

The work was in part supported by the National Key Research and Development Program of China (Grant No.2016YFD0101900), the Defense Industrial Technology Development Program (Grant No. JCKY2018110C165), the Major Technological Innovation Projects in Hubei Province (Grant No. 2019AAA024) and Sanya Science and Education Innovation Park of Wuhan University of Technology (Grant No. 2020KF0054).

REFERENCES

[1] J. J. Roldán, J. del Cerro, D. Garzón-Ramos, P. Garcia-Aunon, M. Garzón, J. de León, and A. Barrientos, "Robots in agriculture: State of art and practical experiences," in *Service Robots*, A. J. R. Neves, Ed. Rijeka: IntechOpen, 2018, ch. 4.

[2] D. Slaughter, D. Giles, and D. Downey, "Autonomous robotic weed control systems: A review," *Computers and Electronics in Agriculture*, vol. 61, no. 1, pp. 63–78, 2008.

[3] R. Mur-Artal and J. D. Tardós, "ORB-SLAM2: an open-source SLAM system for monocular, stereo and RGB-D cameras," *IEEE Transactions on Robotics*, vol. 33, no. 5, pp. 1255–1262, 2017.

[4] T. Qin, S. Cao, J. Pan, and S. Shen, "A general optimization-based framework for global pose estimation with multiple sensors," *arXiv preprint arXiv:1901.03638*, pp. 1–7, 2019.

[5] R. Wang, M. Schwörer, and D. Cremers, "Stereo dso: Large-scale direct sparse visual odometry with stereo cameras," in *IEEE International Conference on Computer Vision*, 2017, pp. 3923–3931.

[6] D. Gálvez-López and J. D. Tardos, "Bags of binary words for fast place recognition in image sequences," *IEEE Transactions on Robotics*, vol. 28, no. 5, pp. 1188–1197, 2012.

[7] T. Pire, M. Mujica, J. Civera, and E. Kofman, "The rosario dataset: Multisensor data for localization and mapping in agricultural environments," *The International Journal of Robotics Research*, vol. 38, no. 6, pp. 633–641, 2019.

[8] M. Imperoli, C. Potena, D. Nardi, G. Grisetti, and A. Pretto, "An effective multi-cue positioning system for agricultural robotics," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 3685–3692, 2018.

[9] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: An efficient alternative to sift or surf," in *International Conference on Computer Vision*, 2011, pp. 2564–2571.

[10] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," in *European Conference on Computer Vision*, 2006, pp. 404–417.

[11] G. Klein and D. Murray, "Parallel tracking and mapping for small ar workspaces," in *IEEE/ACM International Symposium on Mixed and Augmented Reality*, 2007, pp. 225–234.

[12] T. Pire, T. Fischer, J. Civera, P. De Cristóforis, and J. J. Berlles, "Stereo parallel tracking and mapping for robot localization," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2015, pp. 1373–1378.

[13] R. A. Newcombe, S. J. Lovegrove, and A. J. Davison, "Dtam: Dense tracking and mapping in real-time," in *International Conference on Computer Vision*, 2011, pp. 2320–2327.

[14] J. Engel, T. Schöps, and D. Cremers, "Lsd-slam: Large-scale direct monocular slam," in *European Conference on Computer Vision*, 2014, pp. 834–849.

[15] C. Forster, Z. Zhang, M. Gassner, M. Werlberger, and D. Scaramuzza, "Svo: Semidirect visual odometry for monocular and multicamera systems," *IEEE Transactions on Robotics*, vol. 33, no. 2, pp. 249–265, 2017.

[16] J. Engel, V. Koltun, and D. Cremers, "Direct sparse odometry," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 3, pp. 611–625, 2018.

[17] J. Engel, V. Usenko, and D. Cremers, "A photometrically calibrated benchmark for monocular visual odometry," *arXiv preprint arXiv:1607.02555*, pp. 1–9, 2016.

[18] G. Sibley, G. S. Sukhatme, and L. H. Matthies, "The iterated sigma point kalman filter with applications to long range stereo," in *Robotics: Science and Systems*, vol. 8, no. 1, 2006, pp. 235–244.

[19] P. Pinggera, D. Pfeiffer, U. Franke, and R. Mester, "Know your limits: Accuracy of long range stereoscopic object measurements in practice," in *European Conference on Computer Vision*, 2014, pp. 96–111.

[20] P. Ferrara, A. Piva, F. Argenti, J. Kusuno, M. Niccolini, M. Ragaglia, and F. Uccheddu, "Wide-angle and long-range real time pose estimation: A comparison between monocular and stereo vision systems," *Journal of Visual Communication and Image Representation*, vol. 48, pp. 159–168, 2017.

[21] F. Auat Cheein, G. Steiner, G. Perez Paina, and R. Carelli, "Optimized eif-slam algorithm for precision agriculture mapping based on stems detection," *Computers and Electronics in Agriculture*, vol. 78, no. 2, pp. 195–207, 2011.

[22] U. Weiss and P. Biber, "Plant detection and mapping for agricultural robots using a 3d lidar sensor," *Robotics and Autonomous Systems*, vol. 59, no. 5, pp. 265–273, 2011.

[23] M. A. Juman, Y. W. Wong, R. K. Rajkumar, and L. J. Goh, "A novel tree trunk detection method for oil-palm plantation navigation," *Computers and Electronics in Agriculture*, vol. 128, pp. 172–180, 2016.

[24] J. M. Mendes, F. N. dos Santos, N. A. Ferraz, P. M. do Couto, and R. M. dos Santos, "Localization based on natural features detector for steep slope vineyards," *Journal of Intelligent & Robotic Systems*, vol. 93, no. 3-4, pp. 433–446, 2019.

[25] W. Winterhalter, F. V. Fleckenstein, C. Dornhege, and W. Burgard, "Crop row detection on tiny plants with the pattern hough transform," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 3394–3401, 2018.

[26] J. Underwood, A. Wendel, B. Schofield, L. McMurray, and R. Kimber, "Efficient in-field plant phenomics for row-crops with an autonomous ground vehicle," *Journal of Field Robotics*, vol. 34, no. 6, pp. 1061–1083, 2017.

[27] I. Sa, C. Lehnert, A. English, C. McCool, F. Dayoub, B. Upcroft, and T. Perez, "Peduncle detection of sweet pepper for autonomous crop harvesting—combined color and 3-d information," *IEEE Robotics and Automation Letters*, vol. 2, no. 2, pp. 765–772, 2017.

[28] W. Winterhalter, F. V. Fleckenstein, C. Dornhege, and W. Burgard, "Crop row detection on tiny plants with the pattern hough transform," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 3394–3401, 2018.

[29] A. Ahmadi, L. Nardi, N. Chebrolu, and C. Stachniss, "Visual servoing-based navigation for monitoring row-crop fields," in *IEEE International Conference on Robotics and Automation*, 2020, pp. 4920–4926.

[30] J. Dong, J. G. Burnham, B. Boots, G. Rains, and F. Dellaert, "4d crop monitoring: Spatio-temporal reconstruction for agriculture," in *IEEE International Conference on Robotics and Automation*, 2017, pp. 3878–3885.

[31] N. Chebrolu, P. Lottes, T. Läbe, and C. Stachniss, "Robot localization based on aerial images for precision agriculture tasks in crop fields," in *International Conference on Robotics and Automation*, 2019, pp. 1787–1793.

[32] C. Potena, R. Khanna, J. Nieto, R. Siegwart, D. Nardi, and A. Pretto, "Agricolmap: Aerial-ground collaborative 3d mapping for precision farming," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 1085–1092, 2019.

[33] C. Potena, R. Khanna, J. Nieto, D. Nardi, and A. Pretto, "Collaborative uav-ugv environment reconstruction in precision agriculture," in *Proceedings of the IEEE/RSJ IROS Workshop" Vision-based Drones: What's Next*, 2018.

[34] A. Pretto, S. Aravecchia, W. Burgard, N. Chebrolu, C. Dornhege, T. Falck, F. V. Fleckenstein, A. Fontenla, M. Imperoli *et al.*, "Building an aerial-ground robotics system for precision farming: An adaptable solution," *IEEE Robotics & Automation Magazine*, pp. 1–13, 2020.

[35] F. Raverta Capua, S. Sansoni, and M. Moreyra, "Comparative analysis of visual-slam algorithms applied to fruit environments," in *Argentine Conference on Automatic Control*, 11 2018, pp. 1–6.

[36] R. Comelli, T. Pire, and E. Kofman, "Evaluation of visual slam algorithms on agricultural dataset," in *Reunión de trabajo en Procesamiento de la Información y Control*, 2019, pp. 1–6.

[37] K. Sun, K. Mohta, B. Pfrommer, M. Watterson, S. Liu, Y. Mulgaonkar, C. J. Taylor, and V. Kumar, "Robust stereo visual inertial odometry for fast autonomous flight," *IEEE Robotics and Automation Letters*, vol. 3, no. 2, pp. 965–972, 2018.

[38] M. Burri, J. Nikolic, P. Gohl, T. Schneider, J. Rehder, S. Omari, M. W. Achtelik, and R. Siegwart, "The euroc micro aerial vehicle datasets," *The International Journal of Robotics Research*, vol. 35, no. 10, pp. 1157–1163, 2016.

[39] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.

[40] D. Schubert, T. Goll, N. Demmel, V. Usenko, J. Stückler, and D. Cremers, "The tum vi benchmark for evaluating visual-inertial odometry," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2018, pp. 1680–1687.

[41] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. USA: Cambridge University Press, 2003.

[42] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, "A benchmark for the evaluation of rgb-d slam systems," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2012, pp. 573–580.